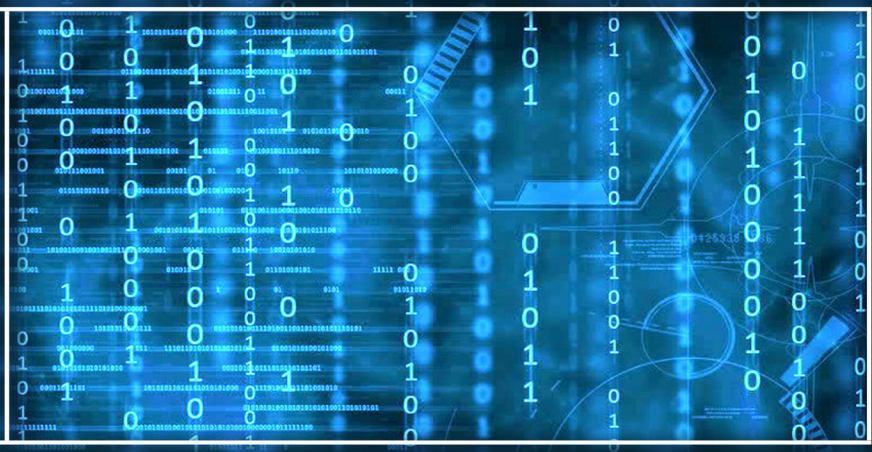
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 13 Issue 8
August 2022



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI
www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Kohei Arai
Editor-in-Chief
IJACSA
Volume 13 Issue 8 August 2022
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Real-Time Wildfire Detection and Alerting with a Novel Machine Learning Approach

A New Systematic Approach of Using Convolutional Neural Network (CNN) to Achieve Higher Accuracy in Automation

Audrey Zhang[1]
Mountain View High School
Mountain View, California, USA

Albert S. Zhang[2]
Amazon
Seattle, Washington, USA

***Abstract*—Up until the end of July 2022, there have been over 38k wildfires in the US alone, decimating over 5.6 million acres. Wildfires significantly contribute to carbon emission, which is the root cause of global warming. Research has shown that artificial intelligence already plays a very important role in wildfire management, from detection to remediation. In this investigation a novel machine learning approach has been defined for spot wildfire detection in real time with high accuracy. The research compared and examined two different Convolutional Neural Network (CNN) approaches. The first approach; a novel machine learning method, a model server framework is used to serve convolutional neural network models trained for daytime and nighttime to validate and feed wildfire images sorted by different times of day. In the second approach that has been covered by existing research, one big CNN model is described as training all wildfire images regardless of daytime or nighttime. With the first approach, a higher detection precision of 98% has been achieved, which is almost 8% higher than the result from the second approach. The novel machine learning approach can be integrated with social media channels and available forest response systems via API's for alerting to create an automated wildfire detection system in real time. This research result can be extended by fine tuning the CNN network model to build wildfire detection systems for different regions and locations. With the rapid development of network coverage such as Starlink and drone surveillance, real time image capturing can be combined with this research to fight the increasing risk of wildfires with real time wildfires detection and alerting in automation.**

Keywords*—*Wildfire detection; CNN (convolutional neural network); machine learning; image processing; model server framework

I. Introduction

The frequency of wildfires is increasing globally, with wildfires occurring this year in unprecedented locations, such as Europe and Yellowstone and Yosemite in the United States. Wildfires cause great property damage and result in numerous injuries and deaths each year. In 2021, there was a record breaking 58,985 wildfires, which ravaged a total of 7.1 million acres [1]. Compared to the 18,229 wildfires and 1.3 million acres lost in 1983 [1], the year when official record-keeping began, this is a sizable increase of 223% [3]. In 2020, California wildfires emitted more than 91 million metric tons of CO_2, that is about 25% of the state's annual fossil fuel emissions and this percentage is forecasted to keep increasing over the next few years [4].

Although wildfires can occur naturally and do provide some beneficial effects like soil nourishment, they need to be controlled in order to mitigate the high levels of CO_2 emission and prevent property loss and casualties. The earlier wildfires can be detected, the better the chance of reducing CO_2 emissions, property damages and life casualties. Fig. 1 breaks down the causes of wildfires [5].

Evidently, from Fig. 1 at least 69% of wildfires stem from human causes, and according to the U.S. Department of Interior, the actual percentage is even closer to 85%. Establishing an automatic wildfire detection and prevention system should be a focal point in reducing the volume of wildfires in the future. With the continuous construction of power lines across the world, a lot of drone investments should be made for surveillance purposes in order to reduce the possibility of man-induced wildfires [7].

To further understand the factors and variables that should be considered for early wildfire detection, In this research paper the top 20 largest California wildfires were examined from information on Inciweb [8], the government's incident information system that displays all present and past cases of wildfires [9]. Wildfires can happen anytime and there is a clear increase in nighttime wildfire intensity due to global warming. Globally, night wildfires have become 7.2% more intense from 2003 to 2020 [10].

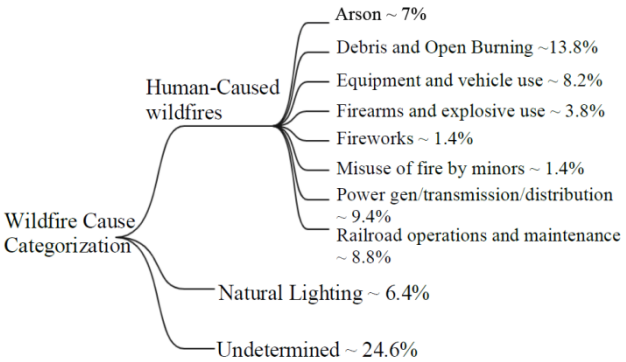


Fig. 1. The Classification of the Causes of Wildfires [5][6].

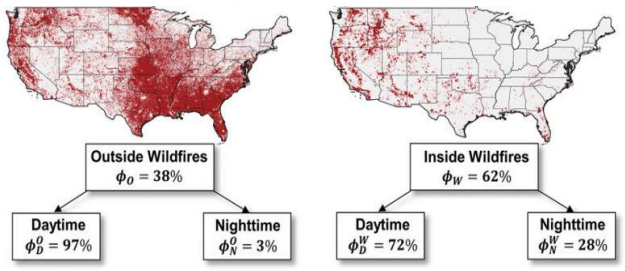


Fig. 2. Moderate Resolution Imaging Spectroradiometer (MODIS) Detected Across the US from 2003 to 2020. Left Represents Outside Wildfires, and Right Represents Inside Wildfires [11].

From the above Fig. 2, ϕ_N and ϕ_D represent the proportion of total fire radiative power detected at nighttime and that detected at daytime respectively. Fig. 2 shows that the increase in nighttime fire activity across the US has been outpacing its daytime counterpart's [11]. So in this research, the images captured at nighttime and the importance of the corresponding network model were important aspects of the research.

The early detection of wildfires is essential to controlling wildfires expeditiously. Academia and industry have come together to solve this increasingly urgent and important problem. Companies, namely soar.earth and pano.ai etc., use the latest technologies such as real time satellite imagery, drone surveillance, and remote network connectivity with Starlink to perform real time image and video capturing, which is a much more accessible and immediate way to detect wildfires. Currently, wildfire spot detection and response times take too long, and in many cases, even a few seconds are important and precious to contain wildfire damages. An automated wildfire detection and alerting system is needed to notify response systems and the public in real time to provide better containment and prevention measures.

This research concluded the convolutional neural network (CNN) paired with the AI model server framework can rapidly and successfully identify wildfires with high accuracy, even source images from daytime versus those from nighttime carry significantly different characteristics. Additionally, the learning models can be extended to different locations and regions with model fine tuning.

II. Background Study

Over the years, many research efforts have delved into the application of artificial intelligence, particularly the use of image recognition and deep learning techniques, to the field of early stage wildfire detection and management. University of California San Diego's WIFIRE center [12] and Piyush and a group of scientists published a review of ML applications to six problem domains: (i) fire detection and mapping; (*ii*) fire weather and climate change; (*iii*) fire occurrence, susceptibility and risk; (*iv*) fire behavior prediction; (*v*) fire effects (*vi*) fire management [2].

Other researchers have explored using classification machine learning models with color features combined with texture classification on superpixel regions of still images [13]. The algorithm uses an RGB color model to detect the color of the fire [13]. Researchers have employed Artificial Neural Networks (ANNs) for fire detection [14] and extended ANNs to wireless sensor networks to create a fire detection system [15]. Various ML methods used in fire detection systems include Support Vector Machines (SVM) to automatically detect wildfires from video frames [16], Genetic Algorithm (GA) for multi-objective optimization of a LiDAR-based fire detection system [17], Bayesian Network (BN) in a vision-based early fire detection system [18], Adaptive Neuro Fuzzy Inference System (NFIS) [19], and K-means Clustering (KM) and fuzzy logic [20].

In the last few years, Academia and industry have come together to find solutions with wildfire detection. Researchers and scientists have found that approaches based on deep convolutional neural networks (CNN) tend to yield the best results for wildfire detection [21]. Tao proposed training CNN models end to end, from raw pixel values to image classifier outputs, and Sharma recommended using imbalanced datasets as inputs to these networks to simulate real life scenarios [22]. In 2019, an adaptive pooling approach of conventional image processing techniques and convolutional neural networks provided even higher accuracy and reliability [23]. Recently a group of researchers at Shanghai University have used CNN and satellite images for wildfires detection [24].

From the above research papers the following conclusions can be drawn:

1) So far, all investigations are based on smoke detection. However, smoke detection using wildfire images taken during the night is not effective, especially with smaller datasets. There has been no research thus far in creating separate learning models for nighttime and daytime wildfires.

2) The wildfire images from nighttime carry large variations in color, texture and shapes. No research papers have talked about those significant variations against smoke based detection.

3) How a CNN Network model can be generic and flexible across various locations and regions for wildfires detection.

4) A self-learning and automated process is imperative to detect wildfires very early.

In this research paper, a novel systematic approach for automatically detecting wildfires and alerting response systems were proposed and implemented with the following three major advances:

1) Google Cloud Platform (GCP) was used in this research to build Convolutional Neural Network (CNN) learning models for daytime and nighttime wildfires.

2) A modern model server architecture to serve the models with input images to achieve high accuracy regardless the time of those images were taken.

3) To make this work more generic so that it can be leveraged at different locations, Convolutional Neural Network Fine tuning is explored to adapt and enhance the network models to have the same high accuracy across locations.

III. Research Methodology

A. Data and Data Preparation

For this research project, a combination of several different sources of wildfire image data were used. The images are taken from Kaggle, open source projects, and Google images. Most of the images have been properly labeled with wildfire and non-wildfire, but a few are unlabeled, so a few of hours work were spent on manually labeling those images for the purposes of this research. In real life, it is worth noting that image labeling can be achieved by crowdsourcing and having data sourcing companies, such as keymakr.ai, scale.ai, etc, provide us with lots of labeled data. The images were split into daytime and nighttime to build two separate learning models with the same setup. The number of images in the training, validation, and test sets used for the wildfire smoke detection model can be seen in Table I. The following pre-processing steps were performed:

1) Combine all images into one big data set.

2) Filter out images of wildfires in black and white and those with questionable smoke and flare.

3) Classify images into daytime and nighttime wildfire sets

4) Ensure all images are properly labeled, especially for the training and test datasets

TABLE I. Total Dataset Split Train, Validation and Test

Model	***# Fires Daytime***	***# Not Fires Daytime***	***# Fire Nighttime***	***# Not Fire Nighttime***	***# Images***
Train	823	1180	760	820	3.6K
Validation	224	380	212	390	1.2K
Test	212	412	198	320	1.1K
Omitted	32	45	63	66	206
Total	591	2717	733	2096	6.1K

Then the following transformations were performed during data loading to improve the performance of the models. The images were resized and cropped to the empirically determined size of 1040 × 1856 pixels to improve training and inference speed. This operation also enables us to evenly divide the image into overlapping 224 × 224 tiles. Finally, normalization of the images to 0.5 mean and 0.5 standard deviation was conducted, as expected by the deep learning package used (pytorch vision).

B. Implementation of CNN and Alerting Mechanism

Pytorch was used as the underlying framework for the Convolutional Neural Network in this research. Fig. 3 illustrates the overall architecture. Using the Fast R-CNN package, based on the idea of running the CNN just once per image and then finding a way to share that computation across ~1000 proposals, each wildfire image was fed once to the underlying CNN and then selective search was run as usual to generate region proposals. Then, each proposal is projected onto the feature maps generated by the CNN. Fast R-CNN offers an exponential increase in terms of speed [25] over traditional CNN's.

In this research, a model server framework was used. Currently, there are multiple implementations of model server for serving purposes, but the original idea of a low latency and high throughput model server came from the research from UC Berkeley Rise Lab Clipper framework [26]. As of now in the market, there are multiple implementations of model server for serving purposes. TorchServe was used as the implementation framework and combined the API from sunrise-sunset.org [27] to get the trigger point for model serving. Fig. 3 illustrates the model server architecture.

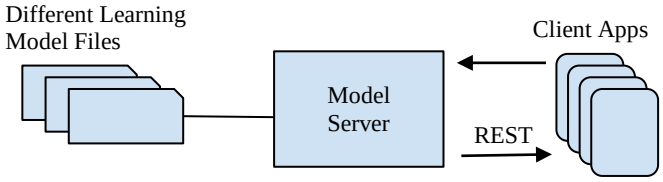


Fig. 3. The Illustration of Model Server Architecture.

The characteristics for wildfires detection in daytime and nighttime can be very different. In the daytime, early detection is heavily based on smoke, so fog, mist, etc. are the primary factors that affect the model's accuracy. At night, early detection focuses more on glares, so lights, fog, etc. are the main factors that could skew accuracy. It is important to treat them differently. After the pre-processing described in section A was completed, three models were trained - (i) one big model with the datasets from both daytime and nighttime combined, (ii) one model for just daytime, and (iii) one model for just nighttime. The models for daytime and nighttime together were aggregated by using the model server framework for the result. The comparison of the results from one big model vs the aggregation of daytime model and nighttime model is discussed in section C. Fig. 4 illustrates the overall systematic approach.

After developing a highly accurate spot wildfire detection system, an alerting and notification mechanism can be established. The higher the detection accuracy is, the lower is the count of false positive alerts. In this investigation, the alerting system was divided into two parts - 1) social channels (Twitter, Facebook, Instagram, etc.) with different social handlers, and 2) SMS text or automated phone calls to police and fire departments - to alert responders to act fast.

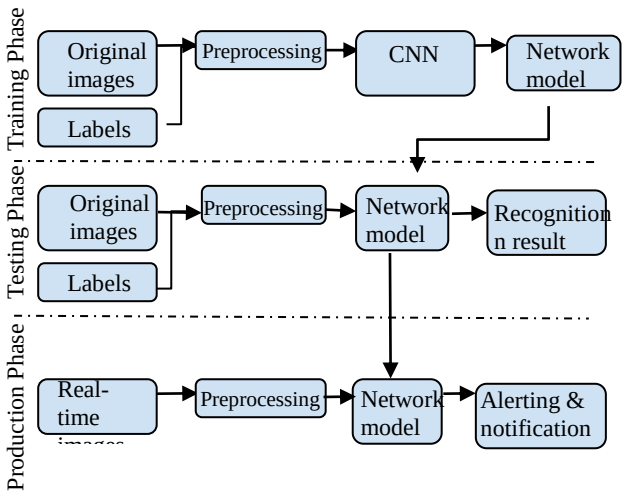


Fig. 4. The Overall Systematic Approach.

To elaborate on the first part, a social handler responsible for monitoring over a specific area can publish wildfire hazard alerts to the people subscribed to it across various social media platforms. This can be very effective in informing people who live around that specific area to prepare themselves for immediate evacuation upon the start of a wildfire. To expand on the second point, cloud services like Twilio can be leveraged to send SMS messages to a given area code for wildfire notifications.

C. Key Metrics and Results

True positive (TP), false positives (FP), true negatives (TN), and false negatives (FN) are calculated between the model predictions and the ground truth labels. For all the experiments, the following two key metrics were used typically for binary classification problems.

- *Precision* - Degree of exactness of the model in identifying only relevant objects.

  It is the ratio of TPs over all detections made by the model, namely: $Precision = \frac{TP}{TP+FP} = \frac{TP}{all\ detections}$

- *Recall* - Measure of the ability of the model to detect all ground-truths instances.

  It is the ratio of TPs over all the ground-truths, namely:

  $Recall = \frac{TP}{TP+FN} = \frac{TP}{all\ ground-truth}$

- Accuracy - Measure the fraction of all instances that are correctly categorized; it is the ratio of the number of correct classifications to the total number of correct or incorrect classifications

  $Accuracy = \frac{TP+TN}{TP+FN+FP+TN} = \frac{TP+TN}{all\ instances}$

A model is said to be a good model if it has high precision and high recall. A perfect model has zero FNs and zero FPs (that is, precision=1 and recall=1). However, it usually is not feasible to attain a perfect model.

In this research, after the model server was employed, the accuracies were calculated by the means of the daytime and nighttime models. Table II shows the comparison of precision, recall and accuracy from different models and approaches at stage training and validation.

TABLE II. RESULTS OF PRECISION, RECALL AND ACCURACY.

Model	***P (%)***	***R (%)***	***A (%)***
Train big model	90.3	71.1	87.5
Train nighttime	97.6	74.3	91.2
Train daytime	98.5	74.7	91.8
Validation big model	89.5	68.3	83.7
Validation nighttime	97.2	73.3	89.7
Validation daytime	98.1	73.9	90.6
Validation with model server	97.9	73.6	89.5

From Table II, it shows that the approach with the model server framework to serve the network models from daytime and nighttime has much higher precision and accuracy of 97.8% and 89.5% vs. one model approach of 89.5% and 83.7% respectively at image validation. Clearly the approach with the model server architecture offers much higher precision and accuracy.

D. Alerting and Notification

In this investigation, a test dataset was used to trigger the process of posting tweets to Twitter. When a test image was fed to the network model, if it is classified as fire, the Twitter API is triggered to generate a message that will be broadcasted to all the subscribers of a Twitter handler automatically in real time. The same mechanism can be implemented to send SMS messages or even automated phone calls.

E. Additional Observations and Future Plans

The results in this research show that spot wildfire can be detected with high precision and accuracy based on the AI Convolutional Neural Network (CNN) learning model with live data stream from monitoring stations, drones or satellites. The cost perspective of the approach from a commercialization perspective will not be discussed in this paper. The computing power for digesting live data and powering the AI model in addition to data storage has a significant cost impact. Soon, real time image capturing will not be an issue in feeding data through a pipeline. Potential edging computing or embedded systems with the AI learning models can be used or deployed to reduce the costs for spot wildfires detection. At the same time, alerting and notification to the public and officials can be established in an automated way. In this research the sunset and sunrise API [27] was used to invoke different models, this may not always be accurate. Wildfire characteristics at dawn and dusk time may get blue, but with more training data, classification accuracies at dawn or dusk will surely increase.

In the future, it would be interesting to do an investigation on the impact of time series images to the complexity of the

network models. Since real life images would be captured via forest monitoring high towers, drones or satellites, by nature they are time series images, this may make the network models even less complex and higher accuracy, but more research and validation are needed. The impact for the time series images to the learning correlations needs further investigation.

In this paper, the dataset used is not very big, the more data that is fed into the CNN, the more powerful the model could become when it gets trained. It would also be an interesting research topic to find a way to automate continuous improvement mechanisms for fine tuning CNN models so that learning models can be generic enough across different locations and regions. Finally, the model server architecture can be further refined and used for different segmentation of conditions with more drastic differences in order to make detection even more robust.

IV. Conclusion

In this research, a novel machine learning convolutional neural network (CNN) with a combination of model server was used to aggregate models for daytime and nighttime to have a higher accuracy. The model server serving different models with different time ranges (approach 1) vs. one big model that did not distinguish daytime and nighttime (approach 2) was compared for accuracy and feasibility. With approach 1, a higher precision of ~98% was achieved vs. approach 2 of ~90%, and with a shorter training time. Approach 1 carries more implementation complexity. This research result shows that wildfires detection accuracy can be improved significantly by considering different models for images from different time intervals and combining them using a model server architecture.

In this research, an alerting and notification system is discussed and can be built to integrate with social media and wildfire responding systems to automate the entire detection and alerting process to have wildfires under control to save lives and reduce property damages.

This investigation is based on the datasets that were collected from the Internet (Kaggle, Google Images and GitHub). In real life, real time videos/images capturing will be more accessible with the development of satellite monitoring, drones and many cases of monitoring towers set up. Those datasets would be more time series and unbalanced data with less variation. The datasets used in this research carry more variations of wildfire scenarios for training the models. A preprocessing with the data sets collected was performed before image feeding to the training process of CNN. With the real life time series images, the preprocessing also could be simplified. The models trained can be used at various places across the globe with CNN fine tuning for spot wildfire detection.

Lastly, the key in wildfire detection, prediction and prevention is to achieve automation, once network models are trained and deployed in the cloud, they can be continuously refined automatically. Live data from real time capturing from various tools can be fed to the models for wildfires detection without much human intervention. A process, feeding of the live data, refining the model and automatically redeploying the model in the cloud will greatly help our societies to fight the increasing risks of wildfires across the globe, and also will help to reduce the carbon emissions resulting from wildfires.

References

[1] "National Fire News | National Interagency Fire Center." [online] Available: https://www.nifc.gov/fire-information/nfn.

[2] P. Jain, S. C. Coogan, S. G. Subramanian, M. Crowley, S. Taylor, and M. D. Flannigan, "A review of machine learning applications in wildfire science and management." Environmental Reviews, vol. 28, no. 4, pp. 478-505, 2020.

[3] "Wildfires and Acres | National Interagency Fire Center." [online] Available: https://www.nifc.gov/fire-information/statistics/wildfires.

[4] E.C. Alberts, "'Off the chart': CO2 from California fires dwarf state's fossil fuel emissions." [online] Available: https://news.mongabay.com/2020/09/off-the-chart-co2-from-california-fires-dwarf-states-fossil-fuel-emissions/.

[5] "Wildfire Causes and Evaluations (U.S. National Park Service)." [online] Available: https://www.nps.gov/articles/wildfire-causes-and-evaluation.htm

[6] "Wildfires in California (I) - Nintil." [online] Available: https://nintil.com/wildfires-california/

[7] M. A. Akhloufi, A. Couturier, and N. A. Castro, "Unmanned Aerial Vehicles for Wildland Fires: Sensing, Perception, Cooperation and Assistance." *Drones*, vol. 5, no. 1, p. 15, 2021

[8] "Top 20 Largest California Wildfires." [online] Available: https://www.fire.ca.gov/media/4jandlhh/top20_acres.pdf.

[9] "Accessibility Friendly List of Incidents" InciWeb - Incident Information System. https://inciweb.nwcg.gov/accessible-view/..

[10] J. K. Balch, "Warming weakens the night-time barrier to global fire." *Nature*, vol. 602, no. 7897, pp. 442-448, 2022.

[11] P. H. Freeborn, W. M. Jolly, M. A. Cochrane, and G. Roberts, "Large wildfire driven increases in nighttime fire activity observed across CONUS from 2003–2020." *Remote Sensing of Environment*, vol. 268, p. 112777-112790, 2022.

[12] "WIFIRE HOME | WIFIRE." [online] Available: https://wifire.ucsd.edu/

[13] D. Y. T. Chino, L. P. S. Avalhais, J. F. Rodrigues, and A. J. M. Traina, "BoWFire: Detection of Fire in Still Images by Integrating Pixel Color and Texture Analysis." 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, 2015.

[14] Sayad Y.O., Mousannif H., and Al Moatassime H. 2019. Predictive modeling of wildfires: a new dataset and machine learning approach. Fire Saf. J. 104: 130–146.

[15] Liu Y., Yang Y., Liu C., and Gu Y. 2015. Forest fire detection using artificial neural network algorithm implemented in wireless sensor networks. ZTE Commun. 13: 12–16.

[16] J. Zhao, Z. Zhang, S. Han, C. Qu, Z. Yuan, and D. Zhang, "SVM based forest fire detection using static and dynamic features." *Computer Science and Information Systems*, vol. 8, no. 3, pp. 821-841, 2011.

[17] A. Cordoba, R. Vilar, A. Lavrov, A. Utkin, and A. Fernandes, "Multi-objective optimisation of lidar parameters for forest-fire detection on the basis of a genetic algorithm." *Optics & Laser Technology*, vol. 36, no. 5, pp. 393-400, 2004.

[18] B. Ko, K.-H. Cheong, and J.-Y. Nam, "Early fire detection algorithm based on irregular patterns of flames and hierarchical Bayesian Networks." *Fire Safety Journal*, vol. 45, no. 4, pp. 262-270, 2010.

[19] K. Angayarkkani and N. Radhakrishnan, "An effective technique to detect forest fire region through ANFIS with spatial data." *2011 3rd International Conference on Electronics Computer Technology*, 2011.

[20] M. BenAmmar and R. Souissi, "A New Approach based on Wireless Sensor Network and Fuzzy Logic for Forest Fire Detection." *International Journal of Computer Applications*, vol. 89, no. 2, pp. 48-55, 2014.

[21] C. Tao, J. Zhang, and P. Wang, "Smoke Detection Based on Deep Convolutional Neural Networks." 2016 International Conference on

Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), 2016.

[22] J. Sharma, O.-C. Granmo, M. Goodwin, and J. T. Fidje, "Deep Convolutional Neural Networks for Fire Detection in Images." Engineering Applications of Neural Networks, pp. 183-193, 2017.

[23] Y. Wang, L. Dang, and J. Ren, "Forest fire image recognition based on convolutional neural network." Journal of Algorithms & Computational Technology, vol. 13, p. 174830261988768, 2019.

[24] Z. Hong, "Active Fire Detection Using a Novel Convolutional Neural Network Based on Himawari-8 Satellite Images." Frontiers in Environmental Science, vol. 10, 2022..

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.

[26] D. Crankshaw and J. Gonzalez, "Prediction-Serving Systems." *Queue*, vol. 16, no. 1, pp. 83-97, 2018

[27] "Sunset and sunrise times API - Sunrise-Sunset.org." https://sunrise-sunset.org/api

Severely Degraded Underwater Image Enhancement with a Wavelet-based Network

Shunsuke Takao, Tsukasa Kita, Taketsugu Hirabayashi
Port and Airport Research Institute,
Department of Infrastructure
Digital Transformation Engineering,
Yokosuka, Japan

***Abstract*—Underwater images are important in marine science and ocean engineering fields owing to giving color information, low cost, and compact. Yet obtained underwater images are often degraded and restoring and enhancing wavelength selective signal attenuation of underwater images depending on complex underwater physical process is essential in practical application. While recently developed deep learning is a promising choice, constructing sufficiently large dataset covering whole real images is challenging, peculiar to underwater image processing. In order to supplement relatively small dataset, previous studies alternatively construct an artificial underwater image dataset based on a physical model or Generative Adversarial Network. Also, incorporating traditional signal processing methods into the network architecture has shown promising success, though enhancement of severely degraded underwater images remains to be a big issue. In this paper, we tackle underwater image enhancement based on an encoder-decoder based deep learning model incorporating discrete wavelet transform and whitening and coloring transform. We also construct a severely degraded real underwater image dataset. The presented model shows excellent results both qualitatively and quantitatively in the artificial and real image dataset. Constructed dataset is available at https://github.com/tkswalk/2022-IJACSA.**

***Keywords*—Underwater image enhancement; deep learning; discrete wavelet transform; whitening and coloring transform**

I. Introduction

Underwater optical images are essential in sensing vast ocean environment. Optical cameras beneficially capture high resolution color information, as well as relatively low cost and compact compared to other acoustic devices. While underwater optical images are essential especially in tasks requiring color information, such as ocean monitoring, maintenance of port facilities, and resource development, serious image degradation is obstacle in efficient utilization. Specifically, wavelength selective color distortion which displays blueish, greenish, and yellowish appearances, or decreased contrast caused by complex underwater physical process worsens the visibility of an underwater image [1], [2], as shown in the upper part of Fig. 1.

To overcome the low visibility of underwater images and expand the scope of application, underwater image enhancement methods based on deep learning have rapidly improved by refining model architecture and training dataset. In underwater image enhancement, deep learning models are mainly trained by mapping degraded images to the corresponding clear images. However, collecting clear and degraded real underwater image pairs is high cost or inherently difficult especially in turbid water in a coastal region. Alternatively, artificial underwater image datasets constructed with a simplified physical model or Generative Adversarial Network (GAN) are employed for training, yet their effectiveness are limited because real underwater images depend on complex physical process and many physical parameters like water body or ambient light and may be apart from artificial images [3], [4]. Subsequently, an artificial underwater image dataset based on the revised underwater image formation model [4] is recently proposed which more reflects real underwater environment [5].

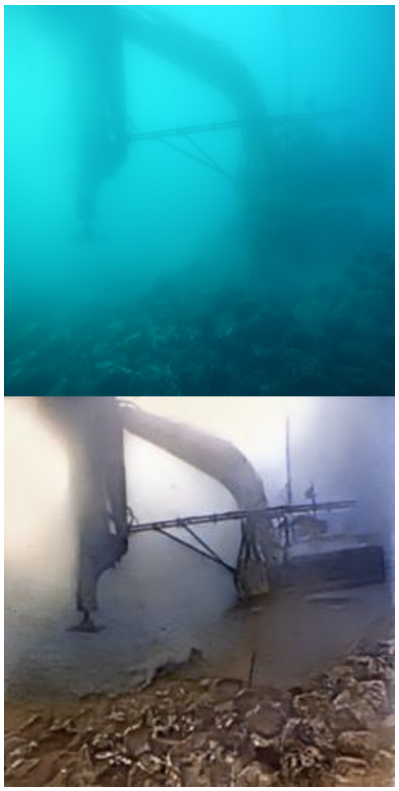


Fig. 1. An Example of a Severely Degraded Underwater Image (Above) and an Enhanced Result by our Model (Below).

Under the constraint of limited amount of data, incorporating traditional signal processing methods into the network architecture is also effective in underwater image enhancement. While the shallow CNN based model incorporating white balance, histogram equalization, and gamma correction has

shown measurable success, enhancement of severely degraded underwater images remains to be a challenging issue [6].

In this paper, we tackle severely degraded underwater image enhancement with an encoder-decoder based network combining discrete wavelet transform and whitening and coloring transform (WCT). The high frequency components of an input image is structurally extracted with discrete wavelet transform in the encoder part, and is preserved by passing them to the decoder part, thereby obtaining a sharp output. Also, as underwater images are quite diverse and display various tones of color and degrees of blurriness, input image features are whitened and mapped to style image features with WCT to stabilize training. The presented model is trained with the recently proposed physically revised artificial underwater image dataset [5] and an elaborated loss function. Also, we present a seriously degraded real underwater image dataset taken in Okinawa, Japan. The constructed dataset includes blueish and greenish images of divers, an underwater construction machine, and rubble mounds of port structures. Our underwater image enhancement model is evaluated with the artificial image dataset and the constructed real image dataset, showing fine results both qualitatively and quantitatively. Our main contributions are summarized as follows:

- We present an underwater image enhancement model combining discrete wavelet transform and whitening and coloring transform.
- We construct a real underwater image dataset including severely degraded blueish or greenish underwater images.
- The presented model successfully removes overall blueish tones of seriously degraded underwater images, mainly outperforming state-of-the-art underwater image enhancement methods both in real and artificial datasets.

II. Related Work

A. Previous Underwater Image Enhancement Methods

Supervised underwater image enhancement models based on Convolutional Neural Network (CNN), Generative Adversarial Network (GAN), and recently appeared Vision Transformer (ViT), have rapidly improved. As models mainly learn pixel transformation tasks, skip connection is often employed not to apart from the original input image. Also, encoder-decoder process is adopted to mitigate the input noise. To be specific, UWCNN is a densely connected CNN model where an input is injected to the different layers with no pooling layers or batch normalization steps [7]. FUnIE-GAN is a fully convolutional conditional GAN model. The generator has five encoder-decoder pairs with several skip connections to enable real time inference [8]. The above two models are either trained with an artificial underwater image dataset. Recently proposed ViT based model is also equipped with several skip connections to stabilize training. To cope with wavelength selective and spatially variant signal attenuation of underwater images, channel-wise attention and spatial-wise attention are incorporated into the architecture [9]. As the difficulty of covering whole real underwater images, incorporating traditional signal processing methods to the network process is effective in underwater image processing. For example, Water-Net is a simple CNN based network which fuses the results of white balance, gamma correction, and histogram equalization [6]. First, three results of each signal processing methods and the original input are fed to the network to predict the three fusion coefficient maps. The predicted three coefficient maps are multiplied by the enhanced results which are obtained by passing through the three independent feature transformation units to reduce the artifacts introduced from the signal processing methods. The refined output is finally obtained by fusing the above three results. Also, discrete wavelet transform is employed to preserve fine image structure [10], [11]. Other than learning based methods, many unsupervised underwater image enhancement methods assume physical model and correct color distortion by imposing white balance, which often requires the estimation of ambient light or average color [12], [13].

B. Previous Underwater Image Datasets

As obtaining sufficient real underwater image pairs is challenging, construction of the dataset itself is important in underwater image processing. Based on a simplified underwater image formation model, [7] constructed an artificially deteriorated underwater image dataset to which visually matches real underwater images by setting the attenuation coefficient to a constant and neglecting other related physical parameters. More recently, based on the revised underwater image formation model [4], an artificial underwater image dataset is proposed which clearly takes into account the dependency of water types, lightning conditions, and camera sensors. The constructed dataset is implied to be more real compared to the previous one [5]. GAN-based approaches generate artificial underwater images by converting initially clear underwater images to degraded ones to cheat the classifier. The model is trained with an unpaired dataset by minimizing Cycle-Consistency loss [8], [14]. Other than artificial underwater images, clearly enhanced real underwater images are collected among results of many conventional enhancement methods by scoring human ranking by hand. This approach is expected to reflect human perceptions, yet is laborious and the sample size is limited to at most a few thousand [6], [9].

III. Methodology

Presented underwater image enhancement model is based on a simple encoder-decoder network architecture with several skip connections, similar to well known U-Net in image segmentation task [15], as shown in Fig. 2. Pooling and upsampling layers are respectively replaced with discrete wavelet transform and inverse discrete wavelet transform to maintain structural information. Whitening and Coloring Transform (WCT) mainly employed in style transfer task is also incorporated into the model to mitigate covariate shift between training data distribution and test data distribution. Brief introduction of discrete wavelet transform and WCT is described followed by the details of model architecture.

A. Signal Reconstruction with Wavelet Transform

The power of discrete wavelet transform (DWT) especially using Haar wavelet is shown in style transfer and inverse problems by generalizing conventional pooling operations

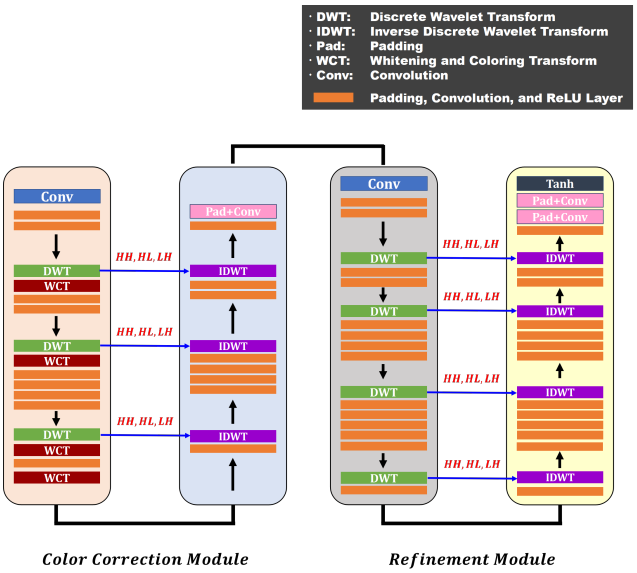


Fig. 2. An Overview of the Model Architecture. An Input Image goes through Several Convolution, Padding, and ReLU Layers Followed by WCT Layers in the Color Correction Module. High Frequency Components, HH, HL, LH, Extracted with Discrete Wavelet Transform in the Encoder are Passed to the Decoder to Preserve Detailed Signal. Compared to Color Correction Module, the Refinement Module is Simply Implemented by Removing the WCT Layers and Adding Several Convolutional Layers to Mitigate Noise.

like average pooling or max pooling, which simply subsamples and summarizes the neighboring pixel information [16], [11], [17]. Haar wavelet operation consists of four kernels, $\{LL^T, LH^T, HL^T, HH^T\}$, where L and H are respectively defined as $L^T := \frac{1}{\sqrt{2}}[1 \quad 1], H^T := \frac{1}{\sqrt{2}}[-1 \quad 1]$. Frequency information is efficiently retained and extracted with L and H, and low frequency signal is captured by L while high frequency signal is captured by H. Inverse discrete wavelet transform (IDWT) is the mirror operation of discrete wavelet transform and is employed for structural reconstruction in the decoder part with minimal noise amplification.

B. Whitening and Coloring Transform (WCT)

The aim of WCT in style transfer is to obtain a stylized image preserving content features [18]. After feature extraction with a pre-trained network, covariance matrix of high dimensional feature maps f_c of content image is first made to be an identity matrix (whitening), followd by singular value decomposition. The whitened content feature $\hat{f}_c$ is then projected onto the eigenspace of the style feature f_s (coloring), described as following procedure:

1) Whitening: Obtain whitened feature $\hat{f}_c = E_c D_c^{-\frac{1}{2}} E_c^T f_c$
2) Coloring: Obtain colored feature $\hat{f}_{cs} = E_s D_s^{\frac{1}{2}} E_s^T \hat{f}_c$

where E_c and D_c are respectively an orthogonal matrix of eigenvectors and a diagonal matrix of the covariance matrix of f_c, and E_s and D_s are that of f_s. Here, $f_c f_c^T = E_c D_c E_c^T$ is satisfied. Colored feature $\hat{f}_{cs}$ satisfies $\hat{f}_{cs}\hat{f}_{cs}^T = f_s f_s^T$, preserving higher order feature correlation which reflects style information. As underwater image distribution is complex and various,

Fig. 3. Our Model Recovers the Severely Degraded Artificial Underwater Image Better (Right) Compared with the Baseline (Left).

TABLE I. PSNR AND SSIM OF OUR MODEL AND BASELINE PER 10 WATER TYPES.

Water-type		I	IA	IB	II	III	1C	3C	5C	7C	9C
PSNR	Propose	16.61	16.316	16.568	15.124	15.8	16.987	15.94	16.214	15.485	15.559
	baseline	16.031	15.514	15.369	13.896	14.351	15.838	14.53	15.124	13.999	14.175
SSIM	Propose	0.702	0.686	0.684	0.608	0.649	0.7	0.652	0.662	0.626	0.629
	baseline	0.684	0.66	0.65	0.566	0.606	0.672	0.609	0.625	0.578	0.582

WCT is incorporated in our model to mitigate the covariate shift between training data and testing data.

C. Network Architecture

The network architecture shown in Fig. 2 is a simple encoder-decoder based model with several skip connections and no pooling layers. In order to preserve detailed image signal, high frequency components extracted with discrete wavelet transform in the encoder part, $\{LH^T, HL^T, HH^T\}$, are passed to the inverse discrete wavelet transform in the decoder part. WCT is incorporated in the color correction module to normalize feature maps.

Input images are first passed through a convolutional layer followed by several convolution, padding, and ReLU activation layers in the color correction module. Then, encoded features go through the discrete wavelet transform layer and low frequency component, LL^T, is processed with WCT and subsequent deeper layers. The remaining high frequency components, LH^T, HL^T, HH^T, are skipped to the decoder part to preserve detailed signal. The encoded features and the passed high frequency components are up-sampled with inverse discrete wavelet transform followed by several convolution, padding, and ReLU activation layers. The subsequent refinement module is similar to the color correction module, but WCT is removed and several convolution, padding, and ReLU activation layers and the last layers of padding, convolution, and hyperbolic tangent activation layer are added to mitigate input noise. Such repeated structure is designed to extract local image structure. Kernel size and stride of all convolutional layers are set to 3 and 1, respectively. We use pre-trained model on photo-realistic style transfer method [11], denoted as baseline, to normalize the complex input distribution caused by the complicated real underwater environment. Our model is similar to [11], but one DWT and IDWT layers, a few convolution, padding, and ReLU activation layers, and the last layers of padding, convolution, and hyperbolic tangent activation layer are added. Here, Fig. 3 shows that

Fig. 4. We Collect Real Underwater Images around Rubble Mounds of Port Structures taken in Okinawa, Japan. The Collected Underwater Images are Significantly Degraded, Showing Blueish or Greenish Appearances.

our model recovers the severely degraded artificial underwater image better (right) compared with the baseline (left). As for quantitative metric, Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), full reference metrics of image quality which reflects human perception, are computed. PSNR and SSIM of our method and the baseline per 10 water types classified by [19], are shown in Table I. Scores of PSNR and SSIM are improved approximately 1 and 0.3, respectively, compared to the baseline in all water types.

D. Loss Function

We combine three loss terms, reconstruction loss L_{rec}, Laplacian pyramid Lap1 loss L_{lap}, and luminance loss L_{lum} between a correct image I_c and an estimated image I_e for training our network, defined as follows:

$$Loss = \alpha L_{rec} + \beta L_{lap} + \lambda L_{lum} \tag{1}$$

where α, β, and λ are hyper parameters.

Reconstruction loss L_{rec} means the pixel wise l1 distance between I_c and I_e, denoted as follows:

$$L_{rec}(I_c, I_e) = |I_c - I_e|_1 \tag{2}$$

Laplacian pyramid Lap1 loss L_{lap} measures differences between I_c and I_e in Laplacian pyramid representation to take into account various frequency components and get a structural image [20], [21], defined as:

$$L_{lap}(I_c, I_e) = \sum_{i}^{2^{2i}} |L^i(I_c) - L^i(I_e)|_1 \tag{3}$$

Here, $L_i(I)$ means the i-th level of the Laplacian pyramid representation of an image I [22].

Also, we propose luminance loss L_{lum}. The luminance loss measures pixel wise difference between I_c and I_e of their luminance components after transforming to YCbCr color space, described as follows:

$$L_{lum} = |Y(I_c) - Y(I_e)|_1 \tag{4}$$

where luminance component Y can be defined as:

$$Y = 0.299R + 0.587G + 0.114B$$

Here, R, G, and B mean the red, green and blue channels of the original image, respectively. The luminance loss is proposed to facilitate training, as luminance components are less susceptible to color tones of an underwater image.

IV. Experiments

A. Construction of Real Underwater Image Dataset

We collected real underwater images around rubble mounds of port structures in Okinawa, Japan. The constructed dataset contains significantly degraded underwater images of an underwater construction machine, a diver, and rubble mounds, which are taken with GoPro HERO4. As shown in Fig. 4, the underwater images are directly taken by a diver or a camera mounted with the upper part of the construction machine, showing blueish or greenish appearances. The constructed dataset is available at https://github.com/tkswalk/2022-IJACSA.

B. Experimental Setting

We train the model with a recently proposed artificial underwater image dataset [5] based on the revised underwater image formation model which more reflects real underwater environment. The model clearly considers the dependencies of related physical parameters, such as water types, lightning conditions, and camera sensors [4]. In the dataset, wavelength data of 10 water types classified by [19], two camera sensors, and the three light spectrum data are employed, namely 60 kinds of artificial images are generated per one image. Clear indoor RGB-D images from NYU Depth Dataset V2 [23] containing depth information are transformed based on the underwater physical model, resulting in 86940 image pairs in total [4], [5]. Among the 1449 original images from NYU Depth Dataset V2, first 1000 images are used for the training data, next 300 images are used for the validation data, and the last 149 images are used for the test data.

A degraded input image is first resized to 256 $\times$ 256 resolutions and mapped to an enhanced image. The coefficients of the loss function, α, β, and λ, are respectively set to 1, 10, 1. Adam optimizer [24] is adopted and the learning rate is set to 0.0001. The training epoch is 80 and the model is implemented with PyTorch and GeForce RTX 2080 Ti GPU.

C. Results and Discussions of Artificial Underwater Images

We qualitatively and quantitatively compare the restoration results with available state-of-the-art underwater image enhancement methods. As shown in Fig. 5, FUnIE-GAN (4th row) [8], UWCNN [7] (7th row), Water-Net [6] (8th row), and U-Transformer [9] (9th row) are evaluated for the deep learning based approaches, while results of retinex-based theory (5th row, denoted as Retinex) [13] and underwater dark channel prior (6th row, denoted as UDCP) [12] are also compared for the unsupervised methods. The first row of Fig. 5 shows the

TABLE II. RESULTS OF PSNR PER 10 WATER TYPES

PSNR	I	IA	IB	II	III	1C	3C	5C	7C	9C
UWCNN	14.74	13.95	13.31	11.03	11.17	11.31	11.3	10.37	10.89	11.11
FUnIE-GAN	17.3	15.8	14.99	13.54	13.45	15.95	13.57	13.86	12.44	12.17
Water-Net	17.02	15.45	14.6	13.73	13.14	16.52	13.24	12.65	12.4	12.5
Retinex	15.86	14.87	13.97	12.87	12.93	14.52	13.01	13.48	12.39	12.49
UDCP	13.38	12.55	11.69	11.03	10.78	12.4	10.81	10.59	10.36	10.54
U-Transformer	16.63	15.15	14.46	13.11	13.09	15.5	13.21	13.54	12.4	12.59
Ours	16.61	16.32	16.57	15.12	15.8	16.99	15.94	16.21	15.48	15.56

TABLE III. RESULTS OF SSIM PER 10 WATER TYPES

SSIM	I	IA	IB	II	III	1C	3C	5C	7C	9C
UWCNN	0.695	0.655	0.622	0.511	0.511	0.541	0.534	0.508	0.462	0.48
FUnIE-GAN	0.684	0.64	0.611	0.55	0.561	0.644	0.564	0.567	0.53	0.528
Water-Net	0.765	0.683	0.655	0.592	0.599	0.706	0.603	0.586	0.57	0.5793
Retinex	0.715	0.677	0.652	0.579	0.603	0.676	0.606	0.624	0.575	0.5791
UDCP	0.658	0.607	0.544	0.479	0.476	0.588	0.479	0.462	0.452	0.469
U-Transformer	0.651	0.617	0.593	0.538	0.549	0.616	0.553	0.565	0.53	0.536
Ours	0.702	0.686	0.684	0.608	0.649	0.7	0.652	0.662	0.626	0.629

employed test data of indoor images from [23], and they are artificially converted based on a underwater image formation model, as shown in the second row. PSNR and SSIM, full reference metrics measuring the image quality, are calculated for the quantitative evaluation.

The artificial underwater image dataset contains various colors and degrees of degradation which reflects water types or lightning conditions [5]. In qualitative evaluation in Fig. 5, many restoration results are not sufficiently well recovered because of the severe image degradation of an input. While our model relatively well restored blueish, greenish, and yellowish artificial underwater images (3rd row), previous methods hardly improve the visibility (4th [8], 6th [12], and 7th row [7]) or insufficiently output whitish images (5th [13], 8th [6], and 9th row [9]). Also, PSNR and SSIM per 10 water types classified by [19] are respectively shown in Table II and Table III. Our model achieves better performance compared to other methods in 9 out of 10 water types. While our model mainly outperforms other methods in almost all water types, output images are sometimes decolored as shown in the 4th column of the 3rd row.

D. Results and Discussions of Real Underwater Images

Next, we proceed to restoration results of real underwater images, as shown in the 1st row of Fig. 6. Real underwater images of 1st to 3rd column come from the constructed dataset collected in Okinawa, Japan, and the remains come from [6] which contains severely degraded underwater images. Our model (2nd row) restores significantly degraded blueish (1st, 2nd, 3rd, 6th column), greenish (4th column), and yellowish (5th column) underwater images. The output images contain less overall blueish tones compared to results of other methods. Among the results of previous methods, Water-Net [6] (7th row) combining white balance, gamma correction, and histogram equalization, are better also in the yellowish and greenish inputs, yet failed to restore the severely degraded input (1st column). The performance of Water-Net is mainly dominated by the signal processing results as Water-Net fuses outputs of them. FUnIE-GAN [8] (3rd row), GAN based model, hardly improves the visibility and adds grid artifact in severely degraded inputs. UWCNN [7] (6th row), CNN based model, introduces color bias as shown in the 4th and 5th column. Vision transformer based U-Transformer

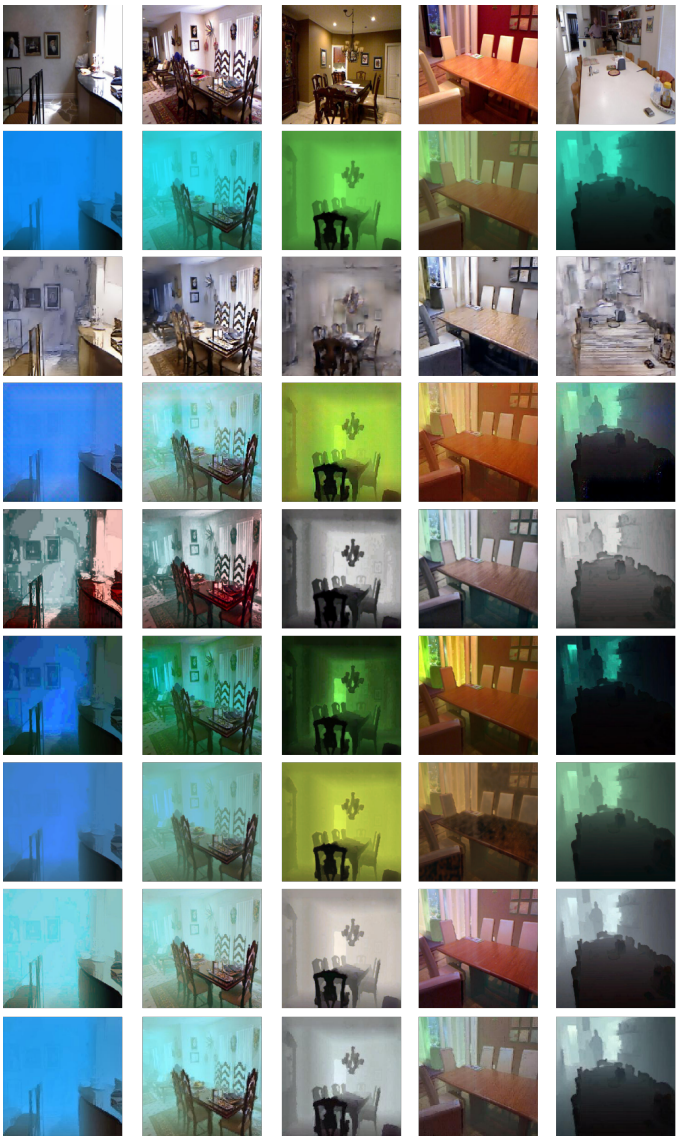


Fig. 5. Restoration Results of Artificial Underwater Images. 1st Row Shows Original Indoor Images, 2nd Row Shows Transformed Input Images, 3rd Row Shows Results of Proposed Model, 4th Row Shows FUnIE-GAN [8]. 5th Row Shows Retinex [13], 6th Row Shows UDCP [12], 7th Row Shows UWCNN [7], 8th Row Shows Water-Net [6], and 9th Row Shows U-Transformer [9].

[9] (8th row) failed to recover greenish and yellowish inputs, respectively shown in the 4th and 5th column. Among the non-learning based methods, Retinex [13] (4th row) corrects a greenish image (4th column), yet also adds reddish color bias in other images (2nd, 3rd, and 6th column). UDCP [12] (5th row), statistical method, hardly improves the overall visibility. As no ground truth is available in real underwater images, PSNR and SSIM are not computed. As real underwater images are tremendously diverse, many supervised models fail to enhance severely degraded underwater images. Among results of previous methods, better results are obtained with Water-Net [6]. Compared to Water-Net trained with a dataset less than 1000 real underwater images, our training dataset is about 100 times larger than that of Water-Net. Also, large amount of severely degraded underwater images are included [5], thus

perceptually better results are tend to be obtained with our model, as shown in the 1st column of Fig. 6.

E. Ablation Study of Loss Function

Ablation study of loss function in Eq. (1) is shown in this section. PSNR and SSIM scores per 10 water types are computed in Table IV. Results of employing only reconstruction loss L_{rec} are denoted as $L1$, plus luminance loss L_{lum} are denoted as $L1 + lum$, and all loss are denoted as ALL. Each loss functions contribute the scores in almost all water types. While proposed luminance loss L_{lum} in Eq. (4) improves less, we observe that the luminance loss stabilizes the training, as it doesn't depend on input color.

TABLE IV. Ablation Study of Proposed Loss Function

Water-type		I	IA	IB	II	III	1C	3C	5C	7C	9C
PSNR	L1	16.37	16.081	16.269	14.832	15.485	16.659	15.62	15.873	15.174	15.265
	L1+lum	16.392	16.1	16.251	14.877	15.551	16.651	15.665	15.857	15.24	15.318
	ALL	16.61	16.316	16.568	15.124	15.8	16.987	15.94	16.214	15.485	15.559
SSIM	L1	0.694	0.675	0.672	0.591	0.633	0.689	0.635	0.646	0.607	0.609
	L1+lum	0.695	0.677	0.673	0.593	0.635	0.69	0.637	0.647	0.609	0.611
	ALL	0.702	0.686	0.684	0.608	0.649	0.7	0.652	0.662	0.626	0.629

V. Conclusion

This study tackles significantly degraded underwater image enhancement with a deep learning model incorporating discrete wavelet transform and whitening and coloring transform. The presented model is trained with the elaborated loss function and recently proposed physically revised artificial underwater image dataset. We also construct real underwater image dataset taken near the rubble mounds of port structures. The dataset characteristically includes severely degraded blueish or greenish underwater images. The presented model outperforms previous state-of-the-art underwater image enhancement models in 9 out of 10 water types in the evaluation employing an artificial underwater image dataset. Also, our model successfully removes blueish tints from real underwater images, showing splendid results qualitatively and quantitatively.

References

[1] Y. Wang, W. Song, G. Fortino, L. Qi, W. Zhang, and A. Liotta, "An experimental-based review of image enhancement and image restoration methods for underwater imaging," *IEEE Access*, vol. 7, pp. 140 233–140 251, 2019.

[2] K. A. Skinner, J. Zhang, E. A. Olson, and M. Johnson-Roberson, "Uwstereonet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7947–7954.

[3] S. Anwar and C. Li, "Diving deeper into underwater image enhancement: A survey," *Signal Processing: Image Communication*, vol. 89, p. 115978, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596520301478

[4] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] S. Takao, "Underwater image sharpening and color correction with a dataset based on correct underwater physical model," in *Submitted*.

[6] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2020.

[7] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognition*, vol. 98, p. 107038, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320319303401

[8] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3227–3234, 2020.

[9] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *arXiv preprint arXiv:2111.11843*, 2021.

[10] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rkhlb8lCZ

[11] A. Jamadandi and U. Mudenagudi, "Exemplar-based underwater image enhancement augmented by wavelet corrected transforms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[12] P. L. J. Drews, E. R. Nascimento, S. S. C. Botelho, and M. F. Montenegro Campos, "Underwater depth estimation and image restoration based on single images," *IEEE Computer Graphics and Applications*, vol. 36, no. 2, pp. 24–35, 2016.

[13] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X. Zhang, and X. Ding, "A retinex-based enhancing approach for single underwater image," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 4572–4576.

[14] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using generative adversarial networks," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7159–7165.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[16] M. Fu, H. Liu, Y. Yu, J. Chen, and K. Wang, "Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 203–212.

[17] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha, "Photorealistic style transfer via wavelet transforms," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[18] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," *Advances in neural information processing systems*, vol. 30, 2017.

[19] N. G. Jerlov, *Marine optics*. Elsevier, 1976.

[20] P. Bojanowski, A. Joulin, D. Lopez-Pas, and A. Szlam, "Optimizing the latent space of generative networks," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 600–609. [Online]. Available: http://proceedings.mlr.press/v80/bojanowski18a.html

[21] H. Wu, J. Liu, Y. Xie, Y. Qu, and L. Ma, "Knowledge transfer dehazing network for nonhomogeneous dehazing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[22] Haibin Ling and K. Okada, "Diffusion distance for histogram comparison," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, 2006, pp. 246–253.

[23] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 746–760.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

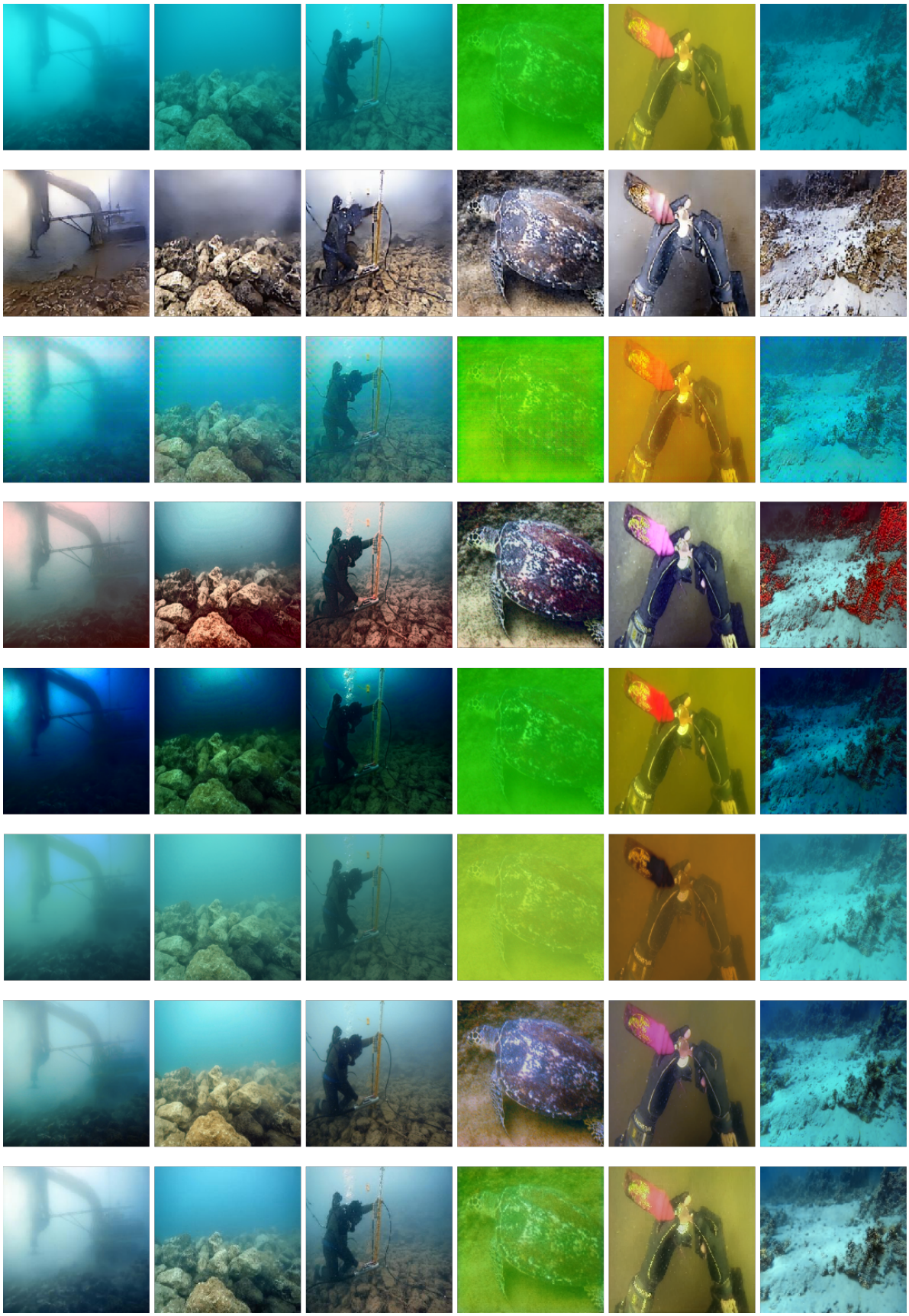


Fig. 6. Restoration Results of Real Underwater Images. 1st Row Shows Input Underwater Images, 2nd Row Shows Proposed Model, 3rd Row Shows FUnIE-GAN [8], 4th Row Shows Retinex [13], 5th Row Shows UDCP [12], 6th Row Shows UWCNN [7], 7th Row Shows Water-Net [6], and 8th Row Shows U-Transformer [9].

Towards Personalized Adaptive Learning in e-Learning Recommender Systems

Massra Sabeima
CSIDS-University of Nouakchott
LIASD, IUT de Montreuil-University Paris8
Montreuil, France

Myriam Lamolle
LIASD, IUT de Montreuil-University Paris8
Montreuil, France

Mohamedade Farouk Nanne
CSIDS-University of Nouakchott
Nouakchott, Mauritanie

***Abstract*—An adaptive e-learning scenario not only allows people to remain motivated and engaged in the learning process, but it also helps them expand their awareness of the courses they are interested in. e-Learning systems in recent years had to adjust with the advancement of the educational situation. Therefore many recommender systems have been presented to design and provide educational resources. However, some of the major aspects of the learning process have not been explored quite enough; for example, the adaptation to each learner. In learning, and in a precise way in the context of the lifelong learning process, adaptability is necessary to provide adequate learning resources and learning paths that suit the learners' characteristics, skills, etc. e-Learning systems should allow the learner to benefit the most from the presented learning resources content taking into account her/his learning experience. The most relevant resources should be recommended matching her/his profile and knowledge background not forgetting the learning goals she/he would like to achieve and the spare time she/he has in order to adjust the learning session with her/his goals whether it is to acquire or reinforce a certain skill. This paper proposes a personalized e-learning system that recommends learning paths adapted to the users profile.**

***Keywords*—e-Learning; adaptive learning; recommendation system; ontology**

I. Introduction

With the broad coverage of the internet, access to learning content through the web has become increasingly easy. A variety of educational systems such as MOOCs[1] [1] have emerged, with an essential mission that is provide educational content, to learners willing to learn; yet the diversity of people implies that each learner has her/his own particular preferences, knowledge and competencies. In that perspective adaptability was a major and essential criteria to add to e-learning systems providing learning resources, to make learning content suitable to learners. This adaptation takes a process that is established in many levels. At the cognitive model level as Ruiz et al. [2] propose, it have to go through the following steps:

- to classify the user by choosing a suitable learning style;
- to present adaptation to system by developing good techniques then conceive that adaptation to suits the user's preferences;
- choosing the right technologies and realization of that adaptation on a computer.

Brusilovsky and Millan [3] on the other hand put focus on the user modeling inside an adaptive system where the user information are a distinctive aspect to consider when the system intervene. The interaction of the user should be noticed with attention, when she/he searches, navigates; but also her/his interest, knowledge, background, learning style, goals, etc. The priority should be given to the content suitable to what user interest in the most. User modeling featured-based or stereotype-based [3] should either way take into consideration the personal information of the individual. A definition of adaptation is the reconfiguration of entities in order to adjust them to a certain request. It can be categorised as the following according to [4]:

- *Machine Centred:* In this case, the learning process is guided by a series of actions from user and analyzed by her/him.
- *User Centred:* The learning resources (lessons) are personalized by learners themselves as stats [4].

Underneath these categories we find several kinds of adaptation [3], [4]. We mention:

- *Adaptation of Content/Adaptive Evaluation:* The content of activities and resources are faced to dynamic change.
- *Adaptation of Visual Presentation:* It represents mainly the components of an interface and their properties, how and where they are displayed.
- *Adaptation of Learning Process:* The learning process is dynamically modified to the manner in which the courses contents are provided in suitable ways.
- *Adaptive Information Filtering:* The system takes care of suitable information retrieval in order to give relevant results to user.
- *Adaptive User Grouping:* This allows a distant learners to collaborate and provide assistance in achieving specific tasks.

However we could not talk about adaptation in a system without mentioning personalization which according to [5] is included into a simple mechanism that need specific technologies to ensure accurate results. Adaptation inside a system

[1]Massive Open Online Courses

takes multiple parameters. The most important one, and regardless the technique used, is the user profile and information; taking into consideration that the user interest and motivation are what will keep her/him continue learning. Therefore well modeling her/his profile is essential. This profile is a personification of different features of user. We note this profile can be built from two types. The first one constitutes a general profile not specific to any user characteristic. The second one is the developed version of the former. After extracting the user information, a personalized profile is created to represent her/him. This study provides a critical overview of previous research related to adaptive recommender systems in e-learning field. Mainly we seek to answer the following research questions:

- What aspect of adaptation should be enhanced and why?
- How is adaptation implemented in recommendation systems?
- How to enhance it?

This paper is organised as follow: Section II will outline similar research work papers. Then in Section III we will underline common techniques and methods used for adaption in adaptive e-learning systems. Section IV presents a comparison about different adaptive systems. Their advantages and drawbacks are highlighted. Section V details our proposition, then in Section VI, we make a position from the current research tendency in e-learning adaptive recommender systems and their techniques, which further consolidates our proposition in the previous section. Finally, Section VII concludes this paper and presents some perspectives.

II. Adaptive E-Learning Systems

Many studies have been conducted in the field of adaptive e-learning systems. In this section, we present an overview of the research papers selected from the literature review. We mention most relative ones from 2017 to 2021. This selection was based on the relevance and the level of the adaptability in e-learning systems in terms of adaptive information filtering, adaptive user profile and adaptive users group. The process of selection was made as follows: among many articles found in Google scholar, ResearchGate[2], DBLP[3] while searching for keywords such as "adaptation in e-learning systems", "recommendation techniques", "adaptative systems", etc. 26 papers were selected based on their relevance, their accuracy, and the number of citations in others research works. Of these 26 articles, we have selected seven (7) to be mentioned, based on the year of publication and direct projection of their content on our research.

Almmouhamadi et al [6] present a survey on the rising techniques of adaptation in educational adaptive systems. They emphasize the two most used techniques of data mining in AI: (i) the predictive which is a prediction of the next tag in general. By selecting a predictor variable or group of variables, these techniques are applied to extract single or multiple variables with predicted values; it is about predicting a missing or unknown item of a dataset, and ii) the descriptive (clustering) one which is based on grouping similar objects. The primary uses of clustering are to segment or categorize (e.g., sorting customer data by age, occupation, and residence) or to extract knowledge in an effort to find subsets of data that are challenging to categorize. This method is about determining a class for an element in a dataset. For instance, we can think of prediction as anticipating the appropriate course of treatment for a certain disease in a specific individual. Whereas the grouping of patients based on their medical records can be considered classification.[4]

George and Lal [7] show how ontology-based recommender systems became an emergent research way in the e-learning field. Those systems address most of issues found in e-learning recommender systems. Giving personalized recommendations to learners is one of the practical applications of employing ontology-based recommender systems. Based on the learner's interests, goals, etc., the recommendations that are given to them become precisely relevant. As a result, the learner is encouraged to finish what she/he started. They illustrated their point of view after a study on research papers published during the last decade concerning the recommender systems in e-learning. They present extraction and modeling techniques used and compare existing recommender systems in e-learning in the scope of these techniques. From another perspective Eke et al. [8] focused on user profiling methods, and the challenges such as multi-dimensional representation, privacy of user's information, cold start problem for new users, temporal behaviour of individuals, limitation of interest, etc. They also discuss the most relevant solution for those challenges such as ontology representation and general purpose profile, and so on.

In [9], Nabizadeh et al. outline the personalization methods and illustrate the challenges facing those methods, and how to improve the existing personalization techniques. Zaoudi et al. [10] present a critical research paper on existing approaches used in learning scenarios and adaptive e-learning situations.

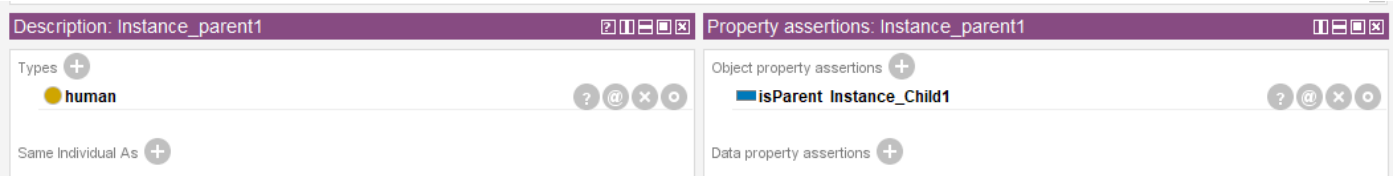
Then Javed et al. [11] present review of a widely used methods in recommender systems, context-based and content-based, and a hybrid method combining multiple methods in order to benefit of the advantage of each method to cover the disadvantages of each one. Just recently, Raj and Ranumol [12] provide a review of research papers on a period of time from 2015 to 2020, with critical study of adaptive recommender systems proposed comparing on one hand methods used in those systems, from the hybrid methods, content or agent-based, semantic web based, etc. On the other hand, they are also comparing the attributes such as the user content rating, learning style, knowledge level, etc.

Table I summarises our comparison of recent research works, which we analyse in Section IV.

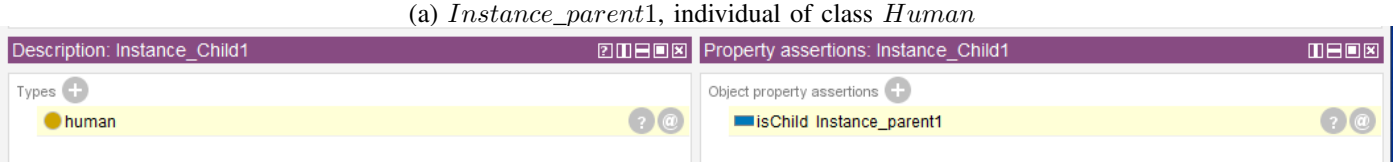
III. Recommendation Techniques

Adaptive recommendation systems can be divided into knowledge-based, content-based, user-based or based on hybrid approaches. They can be categorised according to what the adaptation is based on and on the recommendation techniques

[2]https://www.researchgate.net/
[3]https://dblp.uni-trier.de/
[4]see classification-and-prediction-methods-in-data-mining



(a) $Instance\_parent1$, individual of class $Human$



(b) After inferences

Fig. 1. Example of Inferred Knowledge by the OWL Reasoner HermiT.

used. Following those used techniques the adaptation can be based on:

- *User's Profile:* this method takes into account the characteristics of the user defined by her/his intrinsic characteristics, her/his preferences for the presentation of pedagogical resources to be recommended (text, audio, video, etc.), and the experience of other users with similar profiles. To do this, several techniques have been implemented to model user profile. This allows a learner model to be designed which, according to [13], is the representation of specific characteristics of a learner that may be relevant for a personalized interaction. Managing users' profile allows the user learning style to be predicted. Most of them are based on the widely used learning model "*The Felder and Silverman Learning style Model*" (FSLSM) [14]. It is worth noting that the learner model is not intended to be a representation of the learner's mental state but rather of the learner's characteristics such as personal information (age, gender, country, native language, etc.), cognitive traits, knowledge and skill levels, preferred learning styles, and personal preferences, such as cultural background, format of learning resources (text, audio, video, etc.), preferred language, etc. Munassar and Ali [15] and Aissaoui and Oughdir [16] propose a framework based on user profile modeling using ontologies, which represent the terminology (TBox) and assertions (ABox) such as instances of concept. From a knowledge base, a reasoner checks the consistency of the model and infer new knowledge depending of description logic level used [17].
  For example, let us an ontology O (for didactic purposes):
  $Human(Instance\_parent1)$
  $isParent(Instance\_parent1, Instance\_Child1)$
  $Human.isChild \equiv Human.isParent^{-}$
  We can see below in Fig. 1 how the reasoner HermiT[5] inferred new knowledge.
  Fig. 1(a) shows an individual $Instance\_parent1$ of a class $Human$ with a role (i.e. Object Property assertion in the editor Protégé[6]) $isParent$, which is the inverse role of $isChild$.
  Fig. 1(b) displays, after starting the reasoner HermiT, new knowledge that has been inferred about $Instance\_Child1$ which belongs to class (noted *type* in Fig. 1) $Human$, represented in description logic by $Human(Instance\_Child1)$ and the new object property assertion $isChild(Instance\_Child1, Instance\_parent1)$.
- *Knowledge-Based:* this approach with slight similarity with the user profile based approach, in representing knowledge. It helps making recommendation by extracting information. In general, a reasoning system is behind that decision making, after having well represented knowledge.
  item *Content-Based:* which is based on the content of the started themes. The evaluation of the content is done in an explicit way by the attribution of notes directly to the documents which represent the contents of the topics, or in an implicit way when the system estimates through user interactions the degree of relevance of a document [18].
- *Collaborative Filtering:* which is a widely used method that consists in projecting the preferences of an individual to a group of similar users. In other words, the recommendation is made on the basis of what our neighbors (users with similar profiles) have appreciated [19];
- *Social-Based:* basically these methods can be used to enhance an already existing system, by using social network to create similar groups [20]. One might assume that users who are friends on social networks can have a common interest, or even one user can be interested in a resources because her/his friends were or are interested in taking it. It would be interesting to detect the influencers. User activity on social-media also in recent years formed a good source for recommendation, the time she/he spends watching a video can give an idea on languages she/he understands, and the subjects that interest her/him. Furthermore the content she/he likes and comments or shares also are considerable source. Her/his geolocation history can also be known through her/his publications and the location she/he visits. A variety of information can be extracted through social networks.

[5]www.hermit-reasoner.com/
[6]protege.stanford.edu/

TABLE I. COMPARISON OF ADAPTIVE RECOMMENDER SYSTEMS IN E-LEARNING

Reference	KB	CB	UB	MU
Sarwar et al. [22]	X	-	X	hybrid techniques + ontologies
Agbonifo et al. [23]	-	X	X	collaborative filtrage + ontologies
Vagale et al. [24]	-	X	X	extraction form user model
Boussakssou et al. [25]	X	X	-	Q-learning
Shi et l. [26]	X	X	-	A designed knowledge graph +Bloom's taxonomy
Azzi et al. [27]	X	X	-	Automatic prediction of Learning style + Fuzzy C-means
El Fazazi et al. [28]	X	X	X	MAS+Q-Learning algorithm

- *Hybrid Methods:* they combine two or more approaches of the previous types of recommendation techniques [21]. For example, it can be based on user characteristics by modeling the learner's profile in the first step and, in the second step, recommending resources adaptively with respect to the profile. Usually the combination of these techniques aims to overcome the drawbacks such as the sparsity issue which is due to lack of user rating. Users are reluctant to give feedback on items they have tested. In addition to the sparcity, we mention the cold start problem faced by new users or new item. This issue shows up, when there is not a review/ratings of an item, making it unrecommendable despite its importance and relevancy. The same applies on new users. The difficulty of making recommendations based on a user's profile increases for new users with new profiles.
  Hybride methodes benefits of the advantages of each techniques used. However, some limitations still persist and even new challenges appear in hybrid approaches.

IV. COMPARISON OF ADAPTIVE RECOMMENDER SYSTEM

In this section, we compare adaptive recommender systems proposed firstly in terms of adaptability techniques used: Knowledge-based (KB), Content-Based (CB), User-based (UB) and Method-Used (MU);

and secondly in terms of adaptation itself, based on the two types mentioned in above Section I

Table I gathers recently proposed adaptive e-learning systems using different methods. We notice that [22], [23] are generally using ontology to model content or user profile in their work along side with machine learning techniques, whereas [26], [24], [25] used knowledge designed graphs to represent the user model, or Q-learning [29] with machine learning techniques.

On the other hand, Azzi et al. [27] are more focused on users learning style. They proposes an approach that predicts the user learning style and stores then the collected data. We notice that most of the adaptation techniques used belong to two main ranges, to ontology for modeling and representation, and to machine learning methods. Also other researches tend to combine two at least methods in order to develop an hybrid technique.

These systems mainly evaluate the user performance and make recommendation based on collective preferences like Agbonifo and Akinsete [23] in their work experiments. This method is not the most efficient way due to the lack of specification in those rating (based on what and by who).

Content modeling of learning objects also is another part that should be considered as important as the profile modeling. Moreover it is worth noting the low number of pedagogical resources used in evaluation when seeking for users ratings about a course.

On the other hand the hybrid techniques implemented in those systems whom the semantic part for modeling content and profiles candidate them to be technically efficient as they are using latest technologies tendencies.

Table II gathers previously mentioned systems and lists the types of adaptation presented in the proposed work. We notice that only Boussakssou et al. [25] integrate an adaptation based on user action in their proposed work. Whereas [22], [24] described a group based adaptation and then a content adaptation in [24]. In [28] along with [24] propose an adaptation model that assures course adaptation. This adaptation relies on certain characteristics of the user such as background Knowledge and learning style using Q-learning in El Fazazi et al.s' works [28].

These comparison criteria are selected on the basis of the definition and the different types of adaptation mentioned above.

V. METHODOLOGY

Following our findings on the advantages and disadvantages of current adaptive recommendar systems, we propose a new architecture for piloting and customizing adaptive learning paths, while taking into account the users' profile, the training domain and the available educational resources, and adding synchronization in the collaborative mode between learners wishing to work in collaboration. This system is based on ontologies and a multi-agent system responsible of managing events that occur inside the system. Reasoning on ontologies allows to make tacit information explicit. Among other things, this allows for a better personalization of learning paths. On the other hand, multi-agent systems have shown their great capacity to orchestrate in real time a set of agents.

A. Adaptive and Collaborative Learning Piloting Architecture

This architecture is composed of a multi-agent system (MAS), which contains an agent manager representing the entry point of the main MAPE-K loop of the platform (cf. Fig. 2). This Agent analyzes the requests, processes them and manages the communication between the recommendation agents (RA), responsible for managing the creation of learning paths and the recommendation of educational resources.

TABLE II. COMPARISON IN TERM OF ADAPTATION

Reference	User Based	Machine Based
Sarwar et al. [22]	-	X
Agbonifo et al. [23]	-	-
Vagale et al. [24]	-	X
Boussakssou et al. [25]	X	-
Shi et al. [26]	-	-
Azzi et al. [27]	-	-
El Fazazi et al. [28]	-	X

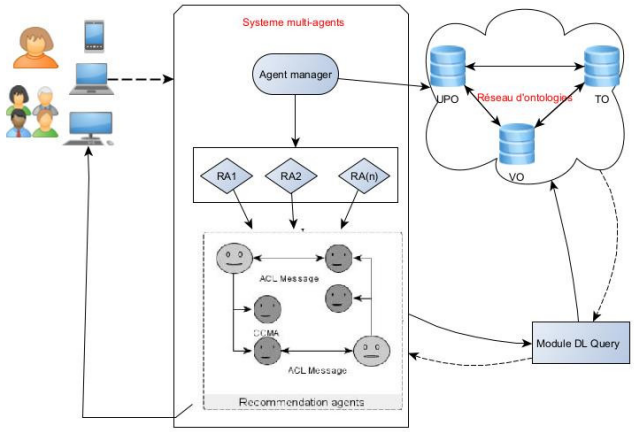


Fig. 2. SPACe-L Architecture.

B. Knowledge Base

The core of the platform is its knowledge representation by a network of ontologies describing three ontological models representing users profiles, training domains and video resources. This modularity is intended to facilitate the interoperability with other ontologies in the respect of the FAIR principles[7]. For example, the user representation (cf. Fig. 3) allows from the ontology FOAF data to be integrated. For the training domain, we can integrate ontologies describing competences like the ontology *COMP2* proposed by G. Paquette [30].

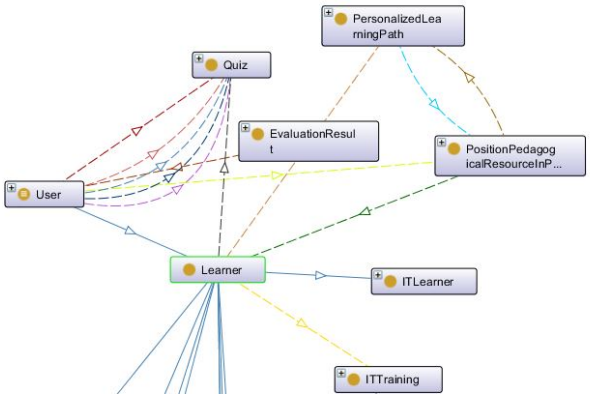


Fig. 3. Partial View of User Profile Ontology (UPO).

The partial view of UPO shown in Fig. 3 contains the personal information and preferences of the users. It mainly describes learners, their specific information, their initial or acquired skills and the personalized learning paths already completed or ongoing.

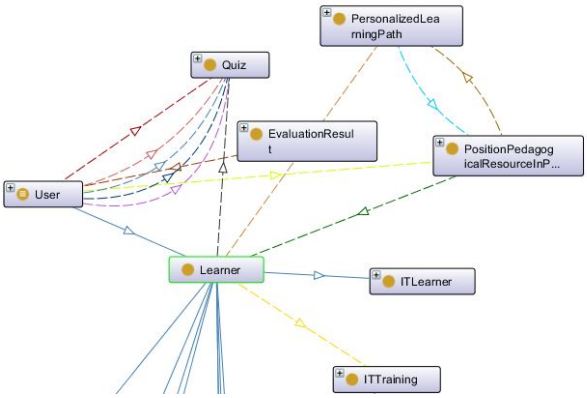


Fig. 4. Partial View of Training Ontology (TO).

The ontology of the training domains (TO) describes the competences to be acquired, the learning objects and the pedagogical resources that can be used in the different pedagogical units (Fig. 4).

C. Multi-Agent System

The multi-agent system (MAS) manages the events that occur in the system, under the supervision of the *Agent manager* who analyzes the requests received and manages the situation according to its nature. It plays a double role according to the learning situation, either individual through a recommendation agent, or collaborative. It then manages a network of recommendation agents for the synchronization of learners. The recommendation agent (RA) is associated to one user or group of users in collaborative situation. The recommendation agent will take care of the generation of the personalized path in the form of a graph by associating relevant pedagogical resources to each node of the graph.

The graph generation is done according to the user profile and the recommendation of the pedagogical resources according to several criteria including the learners' preferences, the duration of the session, but also the qualitative evaluations of the other learners, trainers or experts. The objective here is to maximize a fitness function (cf. Equation 1) which allows to dynamically generate a personalized learning path according to:

$$F = \sum_{i=1}^{n} W_i C_i \quad (1)$$

[7]https://www.go-fair.org/fair-principles/

Where W_i is the weight that defines the importance of C_i and n is the number of selected criteria of a pedagogical resource.

VI. Discussion

Based on the observation form the study done in this paper, it is obvious that adaptation in a recommender systems is essential in order to provide the learner what suits her/him. Therefore, machine learning, ontological and hybrid techniques have been applied in different propositions. We mention in particular the machine learning collaborative filtering technique for its frequent use in those systems. It is reasonable to understand it is widely implicated in most of adaptive e-learning systems since it is a new form of the most traditional method of recommendation (recommendation based on personal user experience). In addition, collaborative filtering system features include recommending an item by classing a list of object based on whether it might be interesting to the user. They include also predicting for a specific item and its rating by a user [31].

It is required, however, to pay attention to some of the challenges that can and have arisen; such as users rating to a certain pedagogical object (courses, learning object, learning path, pedagogical resources, etc.). The integrity of that rating cannot be measured in reality, without exposing publicly the interest of the user or her/his personal information, that leads to another privacy problem. In addition to these problems, we include the unequal number of users and votes on object. George and Lal [7] have pointed out in their research works that the number of users is higher than the number of votes.

Content-based method relies on the interaction of the user and data collection after. Therefore item description is as much as important as the user behaviour, seeing that the recommendation is established based on that. The steps of content-based recommendation techniques are as follows: (i) at the start, item description is stored after analyse, to determine the preferences of a user regarding this item for future use; (ii) then a comparison mechanism is done between user profiles and attributes of these items to sort only related items with similarities with that profile. That said, it still represents multiple drawbacks. Let us mentioning user preferences and interest that change and that affect directly the recommendation. Another issue is the privacy previously mentioned of the user. In order for these methods to have accurate recommendation a large and precise amount of information must be extracted, and that might expose the user privacy policy. Synonymy is also another issue represented by the fact that some of items can have very close description but they still different, which lead to erroneous recommendation [32].

Thus, Semantic Web is a research field existing since the late 90s, especially ontologies, which is promising due to its ability of sharing, reusing and inferring knowledge including through Linked Open Data (LOD), and its level of interoperability. Moreover, it is a good candidate to the FAIR principles (Findable, Accessible, Interoperable, Reusable) [33], [34]. They have also managed to address most of these problems. There is no uniform model for the learner profile or structured material in e-learning, which makes ontology even more relevant [7]. In addition, e-learning requirement can be satisfied by multiple uses of semantic web. This latter one is *Non-linear* as it allows user to describe the situation that she/he is currently in; for example the purpose of her/his learning, and the knowledge acquired. The Semantic Web is also *interactive* that agent can use commonly agreed service language, enabling collaboration between them. Despite the learning resources being distributed on the web, they are linked to one or more commonly accepted ontologies in the scope of semantic web (cf. LOD). Learning materials are distributed on the web, but they are linked to commonly agreed ontologie(s). This allows to build a course that is unique to the user by using semantic querying to find relevant subjects of interest [35]. Application of semantic web can create a responsive learning environment, a personalized learning materials where user only receives what suits her/him, and as much as decentralise content possible.

VII. Conclusions and Future Works

Adaptation in e-learning systems represents a trending research area. In this paper, we presented different adaptive e-learning systems representative of different categories. Several methods were experimented and compared, yet the existing methods have both benefits and drawbacks. The conflict of which one is more effective is still. Mainly the adaptation inside an e-learning environment is user centered even though many researches use the content-based method. Others tend to predict the learning style of the learner, or extract knowledge from user interaction and navigation history; while others lean to use techniques like collaborative filtering and machine learning methods. Except for the fact that user profile modeling remains the main axis to highly adapt content and the learner's need and interest. This being said modeling user is not an easy task to achieve, nor extraction her information by tracing her interactions through the web. Modeling user profile extends representing her interests, competences, expectations of the course and goals. It might reach her mental state at the learning session and after. We highlighted that modelling these new criteria implies a high complexity level in the adaptive process inside an e-learning system. That said the Semantic Web in a side is one sophisticated way to model a profile through the use of ontology. This research field can be highly explored and be employed to improve the current state of adaptive e-learning systems, especially the collaborative learning type which represents an important type of learning and increase learners motivation to reach new competences or reinforce competences.

However some of the relevant questions in that regard still exist such as how can the recommendation systems be improved? And in a more specific manner how is the adequate learning path recommended? How can one be sure we are actually getting the right pedagogical resources? All these questions concerns individual learning situations, it remains those regarding the collaborative situation mentioned earlier, how will the synchronisation between learner be established? Even if established how will the adaptation be maintained? How can we keep learners interested and motivated to finish the training and benefit it the most? How can we integrate citizen science in the scope of collaborative learning? These research questions are important to analyse and to focus on. In this paper we proposed a recommendation system based

on Semantic Web for knowledge representation and multi-agent system that manages the different events in the system. It seems to us that is a good way to answer of the above questions. However there are still several points to improve in order to obtain an advanced adaptive system. starting with enriching our ontology network with ontologies coming from some standards or norms of the educational sciences, and on other side improving response times for learning path and learning resource recommendations during users synchronization from what it concerns the multi-agent system performance.

REFERENCES

[1] M. Zhu, A. R. Sari, and M. M. Lee, "A comprehensive systematic review of mooc research: Research techniques, topics, and trends from 2009 to 2019," *Educational Technology Research and Development*, vol. 68, no. 4, pp. 1685–1710, 2020.

[2] M. d. P. P. Ruiz, M. J. F. Díaz, F. O. Soler, and J. R. P. Pérez, "Adaptation in current e-learning systems," *Computer Standards & Interfaces*, vol. 30, no. 1-2, pp. 62–70, 2008.

[3] P. Brusilovsky and E. Millán, "User models for adaptive hypermedia and adaptive educational systems," in *The adaptive web*. Springer, 2007, pp. 3–53.

[4] A. Klašnja-Milićević, B. Vesin, M. Ivanović, Z. Budimac, and L. C. Jain, "Personalization and adaptation in e-learning systems," in *E-learning systems*. Springer, 2017, pp. 21–25.

[5] S. LaCour, "The future of integration, personalization, and eportfolio technologies," *Innovate: Journal of Online Education*, vol. 1, no. 4, 2005.

[6] K. Almohammadi, H. Hagras, D. Alghazzawi, and G. Aldabbagh, "A survey of artificial intelligence techniques employed for adaptive educational systems within e-learning platforms," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 7, no. 1, pp. 47–64, 2017.

[7] G. George and A. M. Lal, "Review of ontology-based recommender systems in e-learning," *Computers & Education*, vol. 142, p. 103642, 2019.

[8] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke, "A survey of user profiling: State-of-the-art, challenges, and solutions," *IEEE Access*, vol. 7, pp. 144 907–144 924, 2019.

[9] A. H. Nabizadeh, J. P. Leal, H. N. Rafsanjani, and R. R. Shah, "Learning path personalization and recommendation methods: A survey of the state-of-the-art," *Expert Systems with Applications*, vol. 159, p. 113596, 2020.

[10] M. Zaoudi and H. Belhadaoui, "Adaptive e-learning: Adaptation of content according to the continuous evolution of the learner during his training," in *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, 2020, pp. 1–6.

[11] U. Javed, K. Shaukat, I. A. Hameed, F. Iqbal, T. M. Alam, and S. Luo, "A review of content-based and context-based recommendation systems," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 3, pp. 274–306, 2021.

[12] N. S. Raj and V. Renumol, "A systematic literature review on adaptive content recommenders in personalized learning environments from 2015 to 2020," *Journal of Computers in Education*, pp. 1–36, 2021.

[13] A. E. Labib, J. H. Canós, and M. C. Penadés, "On the way to learning style models integration: a learner's characteristics ontology," *Computers in Human Behavior*, vol. 73, pp. 433–445, 2017.

[14] A. Adetunji and A. Ademola, "A proposed architectural model for an automatic adaptive e-learning system based on users learning style," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 5, no. 4, pp. 101–128, 2014.

[15] W. A. Munassar and A. F. Ali, "Semantic web technology and ontology for e-learning environment," *Egyptian Computer Science Journal*, vol. 43, no. 2, pp. 88–100, 2019.

[16] O. E. Aissaoui and L. Oughdir, "A learning style-based ontology matching to enhance learning resources recommendation," in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 2020, pp. 1–7.

[17] F. Baader, "Description logics," in *Reasoning Web: Semantic Technologies for Information Systems,5th International Summer School 2009*, ser. Lecture Notes in Computer Science. Springer–Verlag, 2009, vol. 5689, pp. 1–39.

[18] S. S. Khanal, P. Prasad, A. Alsadoon, and A. Maag, "A systematic review: machine learning based recommendation systems for e-learning," *Education and Information Technologies*, vol. 25, no. 4, pp. 2635–2664, 2020.

[19] M. Ramdane *et al.*, "Les systèmes multi-agents dynamiquement adaptables," Ph.D. dissertation, Université Mentouri Constantine, 2017.

[20] J. K. Tarus, Z. Niu, and G. Mustafa, "Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning," *Artificial intelligence review*, vol. 50, no. 1, pp. 21–48, 2018.

[21] B. Rawat, J. K. Samriya, N. Pandey, and S. C. Wariyal, "A comprehensive study on recommendation systems their issues and future research direction in e-learning domain," *Materials Today: Proceedings*, 2020.

[22] S. Sarwar, Z. U. Qayyum, R. García-Castro, M. Safyan, and R. F. Munir, "Ontology based e-learning framework: A personalized, adaptive and context aware model," *Multimedia Tools and Applications*, vol. 78, no. 24, pp. 34 745–34 771, sep 2019.

[23] O. C. Agbonifo and M. Akinsete, "Development of an ontology-based personalised e-learning recommender system," *International Journal of Computer (IJC)*, vol. 38, no. 1, pp. 102–112, 2020.

[24] V. Vagale, L. Niedrite, and S. Ignatjeva, "Implementation of personalized adaptive e-learning system," *Baltic Journal of Modern Computing*, vol. 8, no. 2, pp. 293–310, 2020.

[25] M. Boussakssou, B. Hssina, and M. Erittali, "Towards an adaptive e-learning system based on q-learning algorithm," *Procedia Computer Science*, vol. 170, pp. 1198–1203, 2020.

[26] D. Shi, T. Wang, H. Xing, and H. Xu, "A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning," *Knowledge-Based Systems*, vol. 195, p. 105618, 2020.

[27] I. Azzi, A. Jeghal, A. Radouane, A. Yahyaouy, and H. Tairi, "A robust classification to predict learning styles in adaptive e-learning systems," *Education and Information Technologies*, vol. 25, no. 1, pp. 437–448, 2020.

[28] H. El Fazazi, M. Elgarej, M. Qbadou, and K. Mansouri, "Design of an adaptive e-learning system based on multi-agent approach and reinforcement learning," *Engineering, Technology & Applied Science Research*, vol. 11, no. 1, pp. 6637–6644, 2021.

[29] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[30] G. Paquette, O. Marino, and R. Bejaoui, "A new competency ontology for learning environments personalization," *Smart Learning Environments*, vol. 8, no. 1, pp. 1–23, 2021.

[31] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The adaptive web*. Springer, 2007, pp. 291–324.

[32] B. Patel, P. Desai, and U. Panchal, "Methods of recommender system: A review," in *2017 international conference on innovations in information, embedded and communication systems (ICIIECS)*. IEEE, 2017, pp. 1–4.

[33] A.-L. Lamprecht, L. Garcia, M. Kuzak, C. Martinez, R. Arcila, E. Martin Del Pico, V. Dominguez Del Angel, S. Van De Sandt, J. Ison, P. A. Martinez *et al.*, "Towards fair principles for research software," *Data Science*, vol. 3, no. 1, pp. 37–59, 2020.

[34] F. Garbuglia, B. Saenen, V. Gaillard, and C. Engelhardt, "D7.5 good practices in fair competence education," Dec. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5785253

[35] B. Dutta, "Semantic web based e-learning," *Documentation Research and Training Centre Indian Statistical Institute, Bangalore*, 2006.

A Comparative Research on Usability and User Experience of User Interface Design Software

Junfeng Wang*, Zhiyu Xu, Xi Wang, Jingjing Lu
College of Design and Innovation, Shenzhen Technology University, Shenzhen, China

***Abstract*—With the development of science and technology, people increasingly rely on intelligent interactive products, thus promoting the vigorous development of the user interface industry. Software with high usability and user experience can improve users' effectiveness and satisfaction, as well as the user viscosity. Taking three design software: Sketch, Adobe XD, and Figma, which is most frequently used by design industry practitioners and students, as research cases, this study compared and discussed the impact of interaction design and interface layout on the usability and user experience combining with subjective experiment methods, scale scoring, user testing and retrospective think-aloud interview, as well as objective experiment method, eye tracking. It is found that the overall usability and user experience of Figma is the best, Adobe XD is the second, and Sketch is the worst. The main reason for this result is that the three software have different degrees of issues in interface layout, information quality, and interaction logic. Based on the results, the optimization suggestions for the usability and user experience of user interface design software are proposed from three perspectives: interface design, information quality and interaction design.**

***Keywords*—Usability; user experience; interaction design; UI design software; eye tracking**

I. Introduction

With the rapid development of the Internet industry and the widespread use of new media, user interface (UI) design emerged, and corresponding tools and software also developed, as well as the application and design of software interfaces have been achieving greater improvement. UI design software play a significant role in the design and development of applications. However, there are few studies focusing on the usability and user experience of UI design software, which does not match the use of such software in the field of UI design [1]. In order to better understand the factors that affect the usability and user experience of this software and to enhance its interaction performance, research on UI design software is necessary. This research compares the overall usability and user experience of Sketch, Adobe XD, and Figma using one objective indicator: eye-movement data, as well as three subjective indicators: scale scores, behavior index and user interviews. It also analyzes the design factors that affect these two aspects. This research contributes to uncovering factors that affect the usability and user experience of UI design software and provides references for optimizing the design and development of such software.

The rest of this paper is divided into 6 sections. The current application of usability and user experience evaluation system in interactive interfaces is discussed Section II. Section III compares the interface layout and interaction design of the research cases. Section IV discusses the preparations for the research. The results from the two dimensions of usability and user experience are analyzed in Section V. Section VI proposes optimization suggestions. And the general conclusion is given in Section VII.

II. Literature Review

Currently, the main observed aspects for assessing the quality of software products are usability and user experience. Usability focuses on system quality and user performance during use, while user experience focuses on the overall satisfaction of users with the system. To analyze and redesign the physical human-computer interface, Ma J et al [2] adopted a subjective and objective multidimensional usability evaluation method. Interaction Experience (IX), a higher-level concept integrating the concepts of usability, user experience as well as accessibility, was proposed by Juergen et al [3] to explore the problems in user-system interaction more precisely.

Vision also is the primary modality for users to interact with the human-computer interface, and related researches show that the fixation metrics [4], saccade metrics [5] and pupil metrics [6] are relative to users' perceived cognitive difficulty and information capturing efficiency. Therefore, combining eye-tracking as an objective physiological assessment method can evaluate the interface quality of software more effectively [7].

For the shopping website pages with different interface layouts and interaction design, Liu C et al [8] acquired the user's subjective perceived usability from the four dimensions of standardization, ease of learning, navigation and attractiveness by questionnaires and interviews. Combined with the eye movement data, they established a relationship model for the perceived usability level of the shopping website pages. Lu C et al [9] used objective eye movement index and behavior index, with subjective scores of four usability indicators, including information clarity, interface comfort, overall satisfaction and performance support, to evaluate the human-machine interaction interfaces with different interface layouts. Pan F [10] proposed an interface usability evaluation model based on eye movement experiment and system usability scale, and conducted quantitative and qualitative analysis on the usability of the ticket purchase website with different information design. Wang Y et al [11] explored the interface layout factors affecting the user experience by making news website pages with different interface layouts as

*Corresponding Author.
The Humanities and Social Sciences Research Planning Fund of the China Ministry of Education.

independent variables, user satisfaction and eye movement index as dependent variables.

The majority of the existing researches focuses on the usability or user experience of information and human-machine interaction interfaces. Most of them explore the user experience from the five dimensions of presentation, framework, structure, scope and strategy, or the three dimensions of instinct, behavior and reflection [12-14]. And the usability is mostly evaluated in terms of ease of learning, effectiveness, satisfaction, efficiency, ease of use [15-17].

Interface design, as one of the most important elements of the software, is influenced by many factors [18]. The interface layout, information design and interaction design are the most significant factors in interface design that can directly affect the usability and user experience. Therefore, this research selects three UI design software with high usage and representative as cases, taking the three significant factors as the entry point to consider their impact on user experience and usability.

III. Comparison of the Interface Layout and Interaction Design

A questionnaire on the use of UI design software was conducted before the experiment, and the usage rate of each common UI design software is shown in Fig. 1. Sketch, Adobe XD and Figma are the three UI design software used most frequently by design industry practitioners and students. The experiment uses the three software as the case study. Interface design is one of the most essential components of software design since the interface is the most direct interaction object for users when using software, and its efficacy and experience have a significant impact on users' intention to use and purchase. Therefore, this research focuses on the core characteristics of interface design: layout and interaction, to compare the similarities and differences of the three software and analyze their impact on usability and user experience.

A. Comparison of Interface Layout

Sketch, Adobe XD, and Figma all adopt the same interface layout, which is shown in Fig. 2. The interface is divided into four areas: the top bar, the left and right sidebars, and the canvas. However, the three software differ in the internal structure of each area, mainly in the following aspects.

The quantity and placement of functions in the top bar vary among the three software, as illustrated in Fig. 3. Both Sketch and Figma include toolbars in the top bar. However, Sketch's top bar contains all tools that users will utilize during the design process, without clear divisions. The top bars of Figma and Adobe XD, on the other hand, are clearly divided into three sections with different functions.

Fig. 4 depicts the layout of the left sidebar in Sketch and Figma. In Sketch and Figma, the left sidebar is the layer list, while in Adobe XD, the left sidebar contains the toolbar and layer list. The property inspector appears on the right sidebar in Sketch, Adobe XD, and Figma, and the interior layout is highly consistent. As for the canvas, there is no distinction among the three software.

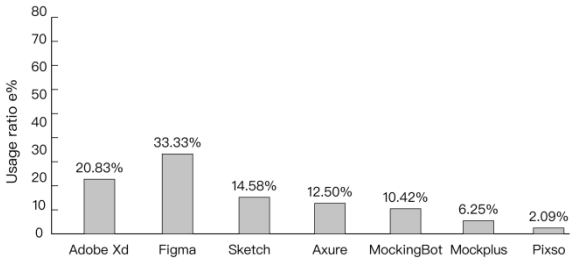


Fig. 1. Usage Rate of UI Design Software.

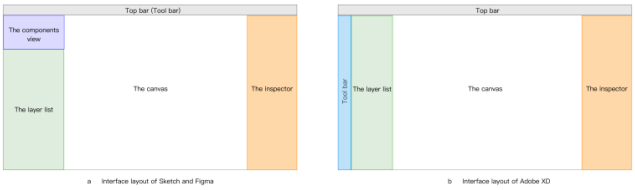


Fig. 2. Interface Layout Comparison.

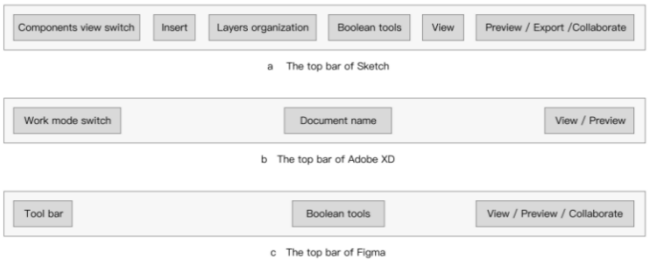


Fig. 3. Comparison of Top Bar Layout.

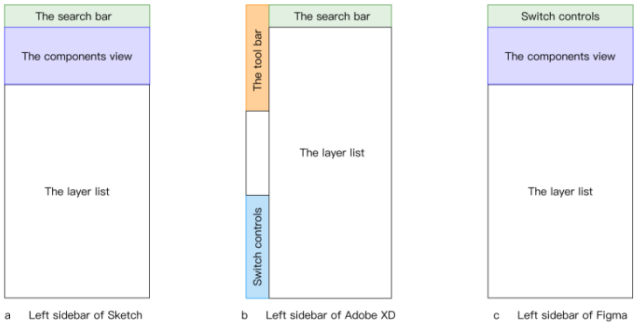


Fig. 4. Comparison of Left Sidebar Layout.

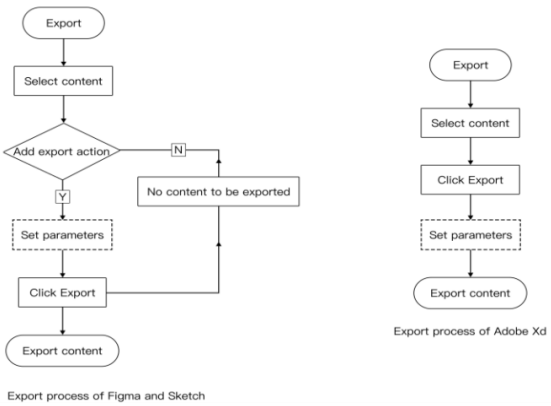


Fig. 5. Comparison of Export Process.

B. Comparison of Interaction Design

In addition to the interface layout, there are some differences and commonalities in the interaction design of the three software. Most of the common functions of the three software interact in the same way, for example, users all add or delete element properties by clicking the Add or Delete controls.

The menu levels of the three software differ. Some functions in Sketch require users to jump to the second level submenu, whereas users only need go to the submenu when using Adobe XD and Figma, and the overall interaction path is shorter. There are also variances in editing property parameters. Figma only shows the parameters frequently edited, with the remainder hidden in the corresponding submenu, while Adobe XD and Sketch show all the parameters in the first level panel of the property inspector.

As demonstrated in Fig. 5, the interaction of export is distinct. When exporting in Adobe XD, users simply select the needed content in the canvas and click Export. In addition, users can choose the content to be exported and set the parameters in the Export dialog. In Sketch and Figma, users must first select the content, then add export action and set the parameters in the property inspector, finally click Export. If users click Export without making the content exportable, the system will notify users that no content is selected or that all frames will be exported by default. This process is more time-consuming than the previous one

IV. Experiment

A. Subjects

Depending on the level of skill, users can be classified as novices, intermediates, and experts. The majority of users are intermediate users, and their amount and frequency of use are consistent. Therefore, the research should primarily focus on intermediate users, and collect their opinions as well as related data on software [19]. The subjects have to be students or practitioners of design industry with at least 6 months of experience in UI design or other related software. Nielsen's research on the number of usability test subjects serves as a basis for the experiment [20]. 17 university students were recruited to participate in this experiment. All subjects were between the ages of 19 and 25, with 9 males and 8 females.

B. Task for Experiment

The key functions of the UI design software were summarized and analyzed before the experiment. A series of interactive tasks covering functions that users use frequently in their daily work were set by combining the results of the UI design software usage questionnaire, and the tasks are shown in Table I.

This experiment chose Sketch 80.1 Chinese version, Adobe XD 45.1.62.364 version, and Figma Chinese v.99.0 as case study, which were the latest versions at the time the experiment was conducted. During the experiment, the three software were run on an iMac computer which has a 27-inch display, and all the subjects were required to finish tasks assigned by the experimenter on this computer.

TABLE I. Experimental Tasks

Task sequence	Task
Task 1	Import [01.jpg] [02.jpg] [03.jpg] into the canvas, and resize them to 200px* 150px
Task 2	Make [Figure1.jpg] [Figure2.jpg] [Figure3.jpg] vertically centered and aligned, keeping their spacing at 30px
Task 3	Add a diameter of 100px circle without fill, whose stroke style is dash and the color number is [666666], thickness of 2; endpoints for round, the dash of 5, gap of 8, transparency of 40%
Task 4	Create the circle drawn in Task 4 as a component and name it [circle].
Task 5	Add the text "Usability Test", with a font size of 14, a font weight of Medium Bold, and a font of Pingfang-SC; set its line height to 18 and the text box to auto width, and adjust the transparency to 50%.
Task 6	Add the component [Shopping Cart] to Frame 1 and detach it from the component.
Task 7	Add the drop shadow effect to [Shopping Cart], set the x-direction parameter to 4, the y-direction to 8, the blur to 8, and the transparency to 3%.
Task 8	Export Frame 1 as 2x size png file to the desktop and name it [test number - name].
Task 9	Save the document to the desktop and rename it to [Test-Number].

C. Evaluation Index System

Combining the definitions of usability and user experience in ISO 9241-11-2018[21], the three dimensions of usability and the primary factors affecting user experience are utilized as the basis for evaluation. The usability and user experience evaluation system based on eye-tracking technology, evaluation scales, retrospective thinking aloud interview, and user testing is constructed from both objective and subjective perspectives, as shown in Fig. 6. By analyzing the subjective and objective data, the elements influencing the usability and user experience of UI design software are summarized [22].

In this experiment, the user testing is a usability evaluation method that collects feedback data on user behavior and satisfaction indicators when using a specific human-machine interface. User testing mainly defines usability problems by observing the process of completing a series of prescribed tasks under a specific scenario and by asking the subjects to record the real usage. In this experiment, the usability evaluation of the software was conducted by testing users, observing and recording the number of failures and the completion time when performing tasks with each of the three software, also using the retrospective thinking aloud interview to obtain participants' experience with the software.

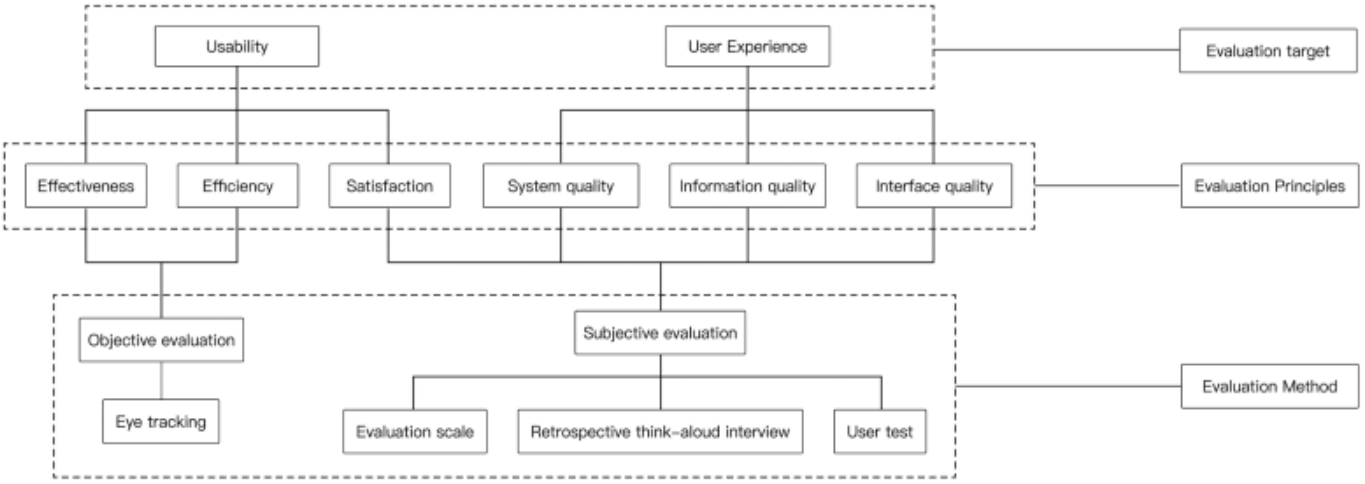


Fig. 6. Usability and User Experience Evaluation System.

The evaluation scales consist of Post-Study System Usability Questionnaire (PSSUQ) and After-Scenario Questionnaire (ASQ). PSSUQ takes information quality, and interface quality as the main evaluation indicators. ASQ quantifies satisfaction by rating the system's support in performing tasks, time spent, and support information. The two questionnaires both adopt a 7-point Likert scale, with 1 indicating "strongly disagree" and 7 indicating "strongly agree", and a higher score on this scale means better usability or satisfaction of the system.

Eye tracking is a common data collection method in usability testing. Based on the usability and user experience evaluation standards, the indicators of task completion time, number of task failures, and total fixation duration in the area of interest (AOI), as well as the heat and gaze plot maps are selected for analysis. The number of failures and task completion time are crucial indicators to evaluate effectiveness and efficiency of the software. The total gaze duration in the area of interest can quantify the difficulty of the object the user is viewing, i.e., the effectiveness of the system's information. The heat map and gaze track map are common visualization forms of eye-tracking data, which mainly demonstrate the user's attention to information and visual search path.

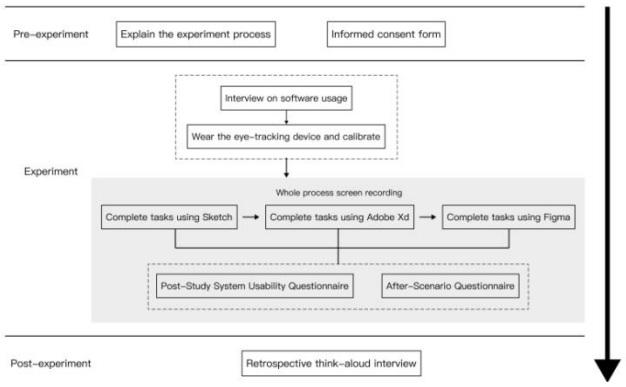


Fig. 7. Experimental Process.

D. Experimental Process

Each subject was given the same task to complete using the three software. Due to the high similarity between the interfaces of Sketch and Figma, the order of subjects using the software was specified as Sketch, Adobe XD, and Figma to prevent learning effects. Only one individual was tested in each experiment in a quiet and bright environment. The procedure of the experiment is shown in Fig. 7.

V. Results

The collected data were assessed for variance chi-square using IBM SPSS.25.0. One-way analysis of variance was used for data with chi-squared variance, while data without chi-squared variance were assessed nonparametrically using the Kruskal-Wallis method. In this experiment, all statistical results were evaluated with the 95% confidence interval. The mean was expressed as "m", while the significance was “p”.

A. Difference in Usability

There are some variances in the degree of support for UI design among the three software, according to the data collected from the experiments.

1) Overall evaluation. The statistical results do not show significant differences when subjects rated the three software overall on the usability scale ($p=0.209>0.05$), but in the post-test interview, nearly 58% of subjects indicated that the overall usability of Figma was significantly better than that of Adobe XD and Sketch. Thus the usability of Figma can be considered higher than Adobe XD and Sketch

2) Effectiveness evaluation. The difficulty of acquiring valid information in an area and the attractiveness of the target to the subject are proportional to the duration of fixation in AOI [23]. Table II shows the duration of fixation in AOI during the completion of tasks using the three software, revealing that the three software do not show significant differences in Task 2, Task 4, and Task 6 ($p_2 = 0.425 > 0.05$, $p_4 = 0.398 > 0.05$, and $p_6 = 0.974 > 0.05$), while the statistics for the other six tasks reveal significant differences.

Furthermore, there are significant differences in the total duration of fixation in AOI ($p = 0.001 < 0.05$). Table III shows the average times of failures while subjects conducting tasks. According to the post-hoc test, the difference between the times of failures using Sketch and Adobe XD is significant ($p=0.038<0.05$). Combining the interface design analysis of the three software, it is found that some ancillary information or icon meaning of Sketch is unclear, which results in subjects spending longer time focusing on the area and increases the difficulty in acquiring valid information compared to Adobe XD and Figma, as confirmed by the post-test retrospective thinking aloud interview. In conclusion, the effectiveness of the three software shows significant differences, and the subjects have difficulty extracting the target information using Sketch, indicating that Sketch has the lowest effectiveness, while the difference in effectiveness between Adobe XD and Figma is not significant.

3) Efficiency evaluation. Table IV depicts results of the average completion time of each task and its one-way ANOVA test. The time of task failed was not included in the statistics [24]. In Table IV, there is no significant difference in the completion time of Task 2, Task 4, Task 5, and Task 6 among the three software, while the total task completion time of Sketch is significantly longer than that of Adobe XD and Figma ($p=0.00<0.05$). The heat map and gaze plot map of Task 3, Task 8 and Task 9, in which subjects performed poorer during the experiment, were selected for comprehensive analysis. Hotspots of different colors can visually reflect the subjects' attention to information and the distribution of gaze points: the longer fixation in red areas, whereas the shorter fixation in green areas [25].

TABLE II. COMPARISON OF DIFFERENCES IN DURATION OF FIXATION IN AOI

Variables	Duration of fixation in AOI (s)			Degree of freedom	Significance
	Sketch	Adobe XD	Figma		
Task 1	111.76	68.08	56.50	2	0.002
Task 2	81.04	56.60	87.52	2	0.425
Task 3	136.48	60.16	64.16	2	0.000
Task 4	53.92	85.81	76.89	2	0.398
Task 5	97.95	66.15	82.00	2	0.036
Task 6	65.31	61.61	61.96	2	0.974
Task 7	98.46	45.75	31.54	2	0.001
Task 8	120.09	45.96	89.48	2	0.013
Task 9	119.47	22.96	45.96	2	0.000
Total duration of fixation	884.48	513.08	576.02	2	0.001

TABLE III. COMPARISON OF DIFFERENCES IN THE TIMES OF TASK FAILURES

Variables	**Times of task failures**			**Degree of freedom**	**Significance**
	Sketch	Adobe XD	Figma		
Times of task failures	1.21	0.21	0.79	2	0.038

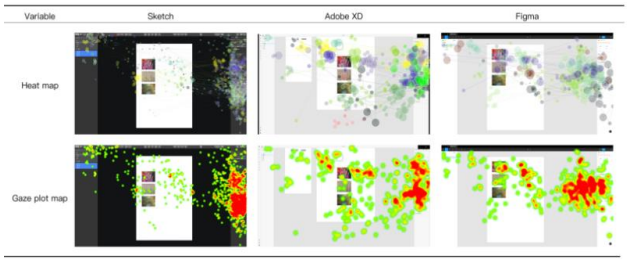


Fig. 8. Eye-movement Diagram when Completing Task 3.

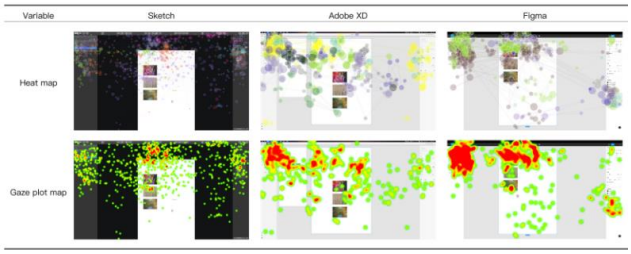


Fig. 9. Eye-movement Diagram when Completing Task 8.

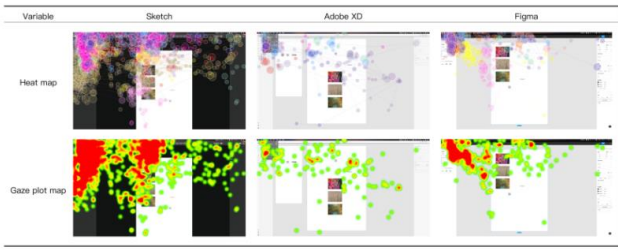


Fig. 10. Eye-movement Diagram when Completing Task 9.

Task 3: Add and edit a shape. In this task, the subjects were required to add a new shape and adjust the parameters of its properties. Fig. 8 shows the specific eye movement map, where hotspots are distributed in Sketch's interface the most widely and the gaze points are the most intensive. Combined with Table II, Table IV and the post-test interview, it is found that Add Layers function of Sketch is hidden in Insert submenu in the top bar, which varies from most of subjects' usage habits. Therefore, subjects were unable to find the required functions quickly. The more amount of eye-track crossings and the longer length of adjacent trajectory segments in Sketch, indicates that subjects' eyes constantly shifted red between the top bar and the canvas, their viewpoints were also more dispersed when using Sketch. In Sketch and Figma, the hotspots area in the right column are larger, which shows that subjects spent more time in this area. In conjunction with the subjects' performance, it reveals that the layout of information in Property Inspector of Sketch and Figma is more compact than Adobe XD, which increases the cognitive burden and leads to lower efficiency of capturing information and completing the task when using the above two software.

Task 8: Export the frame and set the relative parameters. Table IV shows that subjects spent the longest time using Sketch to complete this task and the shortest time using Adobe XD. The specific eye-movement performance is shown in Fig. 9. The subjects' eyes mostly focused on the target function

area when using Adobe XD, and most of them had shorter eye-movement trajectories and fewer red hotspot areas, indicating that their sight did not stay in a specific area for long. Combined with the subjects' performance, it can be seen that since the Export button of Adobe XD is in the Window menu, which is in consistent with majority of subjects' experience, and the parameter information is clear, the subjects could quickly understand the text then completed the task efficiently. The eye-tracking hotspots are dispersed throughout the interface of Sketch, and the amount of eye-tracking crossings is more than the other two software, indicating that the subjects' attention to the interface was scattered. Subjects indicated that Export button in Sketch is small and secluded, and the process is also distinct from their usage habits, making it hard to find the button quickly and requiring them to spend much time trying. Besides, subjects spent the longest total duration of fixation when using Sketch to complete Task 8 (see Table II), implying that the subjects were unable to find the target and acquire information quickly, resulting in low task completion efficiency.

Task 9: Save and rename the file. Combining the statistics in Tables II and IV, it reveals that the subjects' total fixation of duration and completion time when using Sketch are significantly longer than the other two software ($p_1=p_2=0.000<0.05$). Fig. 10 depicts the specific eye-movement diagram for this task. The subjects' eye-movement hotspots when using Sketch concentrate in the menu bar of the window and top bar. The large area of red hotspots and a host of eye-tracking crossings indicate that the subjects' sights shifted between two areas for times, moreover, the browsing speed was slower than Adobe XD and Figma. The subjects claimed in the interview that the main reason for the long time spent on this task using Sketch is that they did not realize the "Copy" meant saving the file as a new file, thus they kept searching for the "Save as" button. It can be seen that ambiguous wording increases the cognitive load of users and reduces their efficiency.

TABLE IV. COMPARISON OF DIFFERENCES IN AVERAGE TASK COMPLETION TIME

Variables	Time to complete each task (s)			Degree of freedom	Significance
	Sketch	Adobe XD	Figma		
Task 1	119.88	71.46	61.26	2	0.001
Task 2	77.52	54.64	61.80	2	0.327
Task 3	120.74	56.00	66.29	2	0.000
Task 4	48.10	78.67	60.65	2	0.273
Task 5	75.68	62.32	74.96	2	0.420
Task 6	59.98	54.80	46.88	2	0.687
Task 7	79.99	48.44	25.53	2	0.001
Task 8	68.11	47.79	38.10	2	0.022
Task 9	106.12	19.89	41.73	2	0.000
Total gaze duration	868.43	549.97	488.16	2	0.000

The analysis of the data leads to the conclusion that the subjects are the most efficient in completing tasks using Figma, followed by Adobe XD, and Sketch is the worst.

4) Satisfaction evaluation. In the post-test interview, the subjects expressed that they preferred Figma to the other software for its succinct interface, reasonable layout, the clear expression of icons, and functions in the toolbar were sorted according to the frequency and priority of use. Additionally, some common parameters are presented in the first-level interface, allowing users to shorten the interaction path and improve efficiency. The majority of subjects thought Sketch's interface was too complicated, which led to a low efficiency in searching for information, and the interaction path of some functions and the interface layout did not conform to usage habits, which hampered the efficiency. On the basis of post-test interview, the satisfaction of the three software can be ranked as Figma>Adobe XD>Sketch.

B. Difference in User Experience

As indicated in Table V, there are significant differences in the overall experience and the information quality of each software.

1) Although there is no statistically significant difference in the overall usability scores of the three software ($p=0.209>0.05$), combined with the specific difference in scores ($m_F=88.17>m_A=80.86>m_S=75.14$) and feedback from subjects in the interview, it can be concluded that there are some differences in the overall user experience among the three software. The data shows that Figma with its succinct interface, clear and organized property control panel, and low learning cost, gives subjects the better experience in UI design.

2) System quality evaluation. Comparing the scores for overall and each subscale of the system quality, it is found that the overall system quality and ($p=0.330>0.05$) do not show significant differences among the three software. Key factors determining the subjects' ratings are whether they can find needed functions quickly and edit parameters efficiently.

The results of the subjects' ratings reveal that Sketch is not very helpful in completing the tasks, primarily due to its complex interface layout, insufficient or ambiguous support information, also the interaction of adding new layers and exporting which differs from usage habits, as evidenced by observations of the experiment and subsequent interviews. While Adobe XD can improve the efficiency, most of subjects indicated that its functions were not as comprehensive as the other two, only supported basic UI design works. In conjunction with results of the scale and interviews, the ranking of system usefulness can be concluded as Figma>Adobe XD>Sketch.

3) Information quality evaluation. The subjects considered Figma to have the best information quality, followed by Adobe XD and finally Sketch. Although there is no significant difference in the overall information quality scores of the three software ($mF=32.36>mA=28.07>mS=26.50$, $p=0.097>0.05$), but in the scores of particular indicators such as information validity ($mF=5.57>mA=4.57>mS=4.14$, $p=0.012<0.05$). Also, subjects indicated that the information in Sketch was confusing and redundant resulting in capturing the needed information inefficiently.

TABLE V. COMPARISON OF POST-TEST SYSTEM USABILITY QUESTIONNAIRE SCORES

Variables		Ratings on the usability of the software			Degree of freedom	Significance
		Sketch	Adobe XD	Figma		
System Usefulness	A1 Ease of use	5.07	5.07	5.57	2	0.591
	A2 Ease of operation	4.64	5.43	5.64	2	0.164
	A3 Efficiency support degree	4.71	5.64	5.57	2	0.151
	A4 System comfort degree	4.50	5.29	5.50	2	0.185
	A5 Ease of Learning	4.71	5.21	5.43	2	0.420
	A6 Performance support	5.14	5.14	5.64	2	0.645
	Total system usefulness	28.79	31.79	33.36	2	0.330
Information Quality	A7 Information guidance	4.00	4.36	4.86	2	0.323
	A8 Fault tolerance	5.07	4.93	5.50	2	0.565
	A9 Information clarity	4.29	4.57	5.43	2	0.102
	A10 Information prominence	4.21	4.71	5.29	2	0.106
	A11 Information validity	4.14	4.57	5.57	2	0.012
	A12 Information structure clarity	4.79	4.93	5.71	2	0.203
	Total information quality	26.50	28.07	32.36	2	0.097
Interface quality	A13 Interface comfort	5.07	5.36	5.93	2	0.280
	A14 Interface preference	5.07	5.50	5.79	2	0.425
	A15 Expectation Satisfaction	4.57	4.79	5.57	2	0.203
	Total interface quality	4.57	15.64	17.29	2	0.251
	A16 Overall satisfaction	5.14	5.36	5.71	2	0.536
	Overall Usability	75.14	80.86	88.71	2	0.209

The main issue with Adobe XD is that the vital information or functions are not prominent enough in the interface, and insufficient auxiliary information to assist users.

4) Interface quality evaluation. The overall and specific indicators of the interface quality only differ in scores, but do not show significant differences. The interface quality of Figma is the best, followed by Adobe XD then Sketch ($m_F > m_A > m_S$). According to the subjects, Figma's interface is the most concise and clear, especially the property inspector, where common properties are shown in the first level panel while less frequently modified properties are collapsed in the second level panel, ensuring a concise interface and shortening the interaction path. Sketch's interface quality is poor for the layout and information architecture of some functions that does not correspond to usage habits; Adobe XD's interface is simple, but some of the frequently edited parameters are folded in the secondary panel, increasing the interaction path. Besides, some frequently used functions are secluded.

VI. DISCUSSION

Based on the results of the preceding analysis, it can be concluded that interface design, information quality, and interaction design highly affect system usability and user experience. Therefore, the designer can optimize the system from the perspective of these three dimensions.

A. Interface Design

Figma's interface is concise, and the color scheme of the interface helps users distinguish the panels clearly. The presentation of the content is clearer than the other two software, and the arrangement of functions is consistent with most users' preferences. The interface of Adobe XD is also concise, but some common tools are not sorted according to the frequency of use. Massive information is presented in Sketch's interface, and the arrangement of some function buttons does not correspond to users' behavior logic. For design software, the interface layout can affect the efficiency and user experience. Different layouts of controls have a significant impact on the eye-movement behavior, so the layout design needs to fully consider the proximity of each area [26]. The interface layout should be reasonably designed based on user's usage habits to make the information more organized [27], reducing the user's cognitive load and effectively improving the efficiency of information acquisition.

B. Information Design

The three software rated low on indicators of information guidance, information prominence, and information clarity in the after-scenario questionnaire. The main issues include unclear semantic expression of icons and some textual information, as well as insufficient auxiliary information. Therefore, when designing icons, it's crucial to ensure that icons are easily recognized and remembered by users [28], and the communication barriers between users and interface can be eliminated [29]. Simultaneously, it is vital to distinguish primary and secondary information in the interface, as well as to emphasize the main information to make it explicit, so as to improve users' cognitive efficiency. The UI design software should also increase or optimize the information that can assist users to work more efficiently. Furthermore, shared resources, such as plug-in libraries and design materials, can boost users' satisfaction.

C. Interaction Design

Efficiency and effectiveness are the main factors that influence users' perceptions of software usability and user experience. Figma outperforms Sketch and Adobe XD in both efficiency and efficacy, according to subjective evaluation and objective data analysis. In general, the interaction path in Figma is more in line with the logic of the user's behavior and functions, allowing users to adapt to the system in a short time and increase efficiency. Specifically, Figma distinguishes

information by frequency of editing, with necessary parameters and function buttons in the first level panel and unnecessary information in the second level panel, to keep operation logic clear and progressive while shortening the interaction path. Consequently when designing such software, the interaction of the system should be concise to make users feel natural and smooth during the process of operation [30]. On this basis, the interaction of the software needs be tailored to the users' preferences, the migration and reuse costs of the software should also be as low as possible.

VII. Conclusion

Three representative and widely used UI design software are chosen as research cases in view of the lack of researches related to usability and user experience of UI design software at this stage. In this paper, effectiveness, efficiency and satisfaction are selected as the evaluation criteria of usability, while system quality, information quality and interface quality are the evaluation criteria of user experience. Objective indicators, the eye movement index and behavioral index, and subjective indicators: evaluation scales, retrospective interviews, and user testing, are used to evaluate and analyze the user experience and usability of the three software. According to the results, each software has varying degrees of problems in interaction design, information quality, and interface layout. Therefore, interface layout, interaction logic, specific interaction pattern, and information visualization should be optimized based on the user's deep needs, as well as the behavior logic, cognitive load, and function logic, so that users can achieve high efficiency through a reasonable human-computer interface.

Though the experiment was designed to prevent learning effects, the results were influenced by the fact that the subjects became familiar with tasks for the same task was performed with three software during the experiment. The number of experiment sample should be expanded in future studies, while experiments should be conducted through a more rigorous form of group control to reduce the interference with the results.

Acknowledgment

This work was supported by the Humanities and Social Sciences Research Planning Fund of the China Ministry of Education. This work was supported by project "Research on age-appropriate interaction design of intelligent voice products" Grant Number 21YJC76007.

References

[1] Han H and Li Y, “Interaction design of mobile learning platforms: a comparative study of usability and user experience”, Modern Intelligence, vol. 41, no. 04, pp. 55-68, 2021.

[2] Ma J and Tan X, “Study of automotive HMI usability evaluation system”, Shanghai Automotive, no. 02, pp.16-19, 2014.

[3] J. Sauer, A. Sonderegger, and S. Schmutz, “Usability, user experience and accessibility: towards an integrative model”, Ergonomics, vol.63, no. 10, pp.1207-1220, Oct. 2020.

[4] J. Gwizdka, R. Hosseini, M. Cole, and S. Wang, “Temporal dynamics of eye-tracking and EEG during reading and relevance decisions”, J. Assoc. Inf. Sci. Technol., vol. 68, no. 10, pp. 2299–2312, Oct. 2017.

[5] R. P. Nazareth and J. H. Kim, “The Impact of Eye Tracking Technology”, in Advances in Usability, User Experience, Wearable and Assistive Technology, Cham, 2020, pp. 524-530. doi: 10.1007/978-3-030-51828-8_69.

[6] Jiang M, Li Q, Wang D, and Xu X, “Usability study of user interface design elements of radiotherapy software based on eye movements and physiological signals”, Packaging Engineering, vol.43, no. 04, pp.163-168, 2022.

[7] Liu Q, Xue C, and F. Hoehn, “Interface usability evaluation based on eye-tracking technology”, Journal of Southeast University (Natural Science Edition), vol.40, no. 02, pp.331-334, 2010.

[8] Liu C, Guo F, Liu W, and Wei Q, “Evaluation and prediction of the elderly’s perceived usability of shopping website homepage interface”, Industrial Engineering and Management, vol. 23, no. 06, pp.101-107, 2018.

[9] Lu C, Hu M, Tang Z, and Zhu Z, “Usability evaluation and redesign of treadmill human-machine interface”, Packaging Engineering, vol.38, no. 22, pp.1-5, 2017.

[10] Pan F, Jiang K, and Wang D, “Usability Design of Ticket Purchase Website Based on Eye Tracking”, Packaging Engineering, vol.41, no. 24, pp. 243–247, 2020.

[11] Wang Y, Wang R, Chen N, Ye J, and Liu W, “News Web Page Design Optimization Based on Eye Tracking Technology”, Technology Innovation and Application, no. 03, pp. 45–50, 2021.

[12] Xu Y, “Research on User Experience of Mobile Game Interactive Interface —— Aimed at Female Users”, Design, vol.32, no. 17, pp. 44-46, 2019.

[13] Xu T, “User Experience of Interactive Interface Design”, Computer Products and Distribution, no. 08, pp. 65, 2018.

[14] Yuan H, Lao C, Zhang Q, Liang B, and Zang C, “Design of Industrial Servo Press Touch-type Interactive Interface Based on User Experience”, Packaging Engineering, vol.40, no. 12, pp.229-235, 2019.

[15] Liang H, and Zhao T, “Evaluation Research on Human-machine Interface Usability of Rope Saw Controller”, Journal of Machine Design, vol.38, no. 06, pp.133-138, 2021.

[16] Zhou W, Xiao D, and Gong M, “Usability of Interactive Interface of Electronic Medical Records”, Packaging Engineering, vol.39, no. 20, pp.248-252, 2018.

[17] Zhao X, Ding Y, Hou W, and Chen X, “On the Usability Evaluation Index System of Complex Information System Interface”, Journal of Graphics, vol.39, no. 04, pp.716-722, 2018.

[18] Tang P, and Li J, “Usability of News APP Interface Based on Eye-tracker Technology”, Packaging Engineering, vol.40, no. 14, pp.247-252, 2019.

[19] Wang M, “Study on the methods offering help to the beginner and the intermediate user”, Packaging Engineering, vol.32, no. 06, pp.85-86+102, 2011.

[20] J. Nielsen and T. Landauer, “A mathematical model of the finding of usability problems”, SIGCHI Conference on Human factors in computing systems, 1993, pp. 206-213.

[21] Ergonomics of Human-system Interaction — Part 11: Usability: Definitions and Concepts, ISO 9241-11:2018.

[22] Liang H and Zhao T, “Evaluation research on human-machine interface usability of rope saw controller”, Machine Design, vol.38, no. 06, pp.133-138, 2021.

[23] Yuan Y, Wu C and Wan X, “Eye-tracking experimental research on consumer online shopping search efficiency”, Packaging Engineering, vol.42, no. 16, pp.218-218+230, 2021.

[24] Li J, Wang J, Hao P, “Research on Usability of the Toolbar of Teaching Interactive Smart Tablet”, J. Phys. Conf. Ser., vol. 1948, no. 1, p. 012150, Jun. 2021.

[25] Wang M, Li Y, Song W, Huang M, and Chen L, “Research on man-machine interface design based on scientific experiments—a case study of the improved design of Gree air conditioner interface”, Journal of Graphology, vol.40, no. 01, pp.181-185, 2019.

[26] J. Holsanova, N. Holmberg, and K. Holmqvist, “Reading information graphics: the role of spatial contiguity and dual attentional guidance”, Applied Cognitive Psychology, vol.23, no.9, pp. 1215-1226, Dec. 2009, doi: 10.1002/acp.1525.

[27] Zhang R, Wang M, Guo Y, Gao S, and Wang D, "Design and evaluation of man-machine interface of oil field water injection system platform", Mechanical design, vol.38, no.07, pp. 118-125, 2021

[28] S. Wiedenbeck, "The use of icons and labels in an end user application program: an empirical study of learning and retention", Behaviour and Information Technology, vol.18, no.02, pp. 68-82, Apr. 1999.

[29] Zhou Y, Luo S, and Chen G, "Icon design based on design semiotics", Journal of Computer-Aided Design and Graphics, vol.24, no.10, pp. 1319-1328, 2012.

[30] Zhang Y, Sun G, Ma X, Xing F, and Qu J, "Optimal design of navigation control interface for deep-sea manned submersible based on human factors", Ship Science and Technology, vol.44, no.01, pp. 154-158, 2022

Experimental Evaluation of Basic Similarity Measures and their Application in Visual Information Retrieval

Miroslav Marinov, Yordan Kalmukov, Irena Valova
Computer Systems and Technologies
University of Ruse
Ruse, Bulgaria

***Abstract*—Searching for similar images is an important feature for image databases and decision support systems in various subject domains. However, it is essential that search results are sorted by degree of similarity in reverse order. This paper presents a comparative analysis of four existing similarity measures and experimentally tests whether they could be used to calculate similarity between images. Metrics could be evaluated by comparing their results to the cumulative human perception of similarity between the same images, obtained by real people. However, this introduces a lot of subjectivism due to non-uniform judgement and evaluation scales. The paper presents a more objective approach - checks which measure performs best in retrieving more images, containing objects of the same type. Results show all four measures could be used to calculate similarity between images, but Jaccard's index performs best in most cases, because it compares features vectors positionally and thus indirectly consider shape, position, orientation and other features.**

Keywords*—Content Based Image Retrieval (CBIR); image search and ranking; similarity measures; image databases

I. Introduction and Related Work

With the development of the Internet and information technology, it has become possible to store and process larger volumes of data, with more and more data in the form of images. This is the basis of the great interest in the approaches and algorithms for image organization, search and retrieval. Naturally, storing large volumes of images requires a new and efficient image retrieval approach. There are two main approaches to image organization and storage - text descriptions, keywords or labels (known as text-based image retrieval) [1] and content based image retrieval [2], [3]. The use of text descriptions is a slow and time consuming process (because of the need a person to describe images with text, not from a computational point of view), so algorithms for content based image retrieval are of greater scientific interest. The main characteristics used in these algorithms are color [4], [5], shape [6], texture [7], spatial features and their combinations [8]. Color is one of the most basic and at the same time distinctive features that hardly changes when you rotate, reduce or increase the size or when changing the orientation of the images. Therefore, the use of color or the color distribution in images at CBIR is the most popular approach among researchers, and yet it is not exhausted and is still subject of interest.

The typical architecture of CBIR systems consists of two main elements. The first is related to the feature extraction of the images and their storage, organization and indexing. The second concerns the assessment of the similarity between the query image and the images in the database. What similarity measures to use and how to assess their suitability for the specific application?

One of the major problems with assessing the visual similarity of images is that there is no classification to use as a criterion. Therefore, it is not possible to make an accurate assessment of the results of the application of the various methods for assessing the similarity of images. It is not possible to use user evaluation (through surveys or any other methods) as the subjective factor in the evaluation is too important and there are undoubtedly huge differences in similarity ratings made by different people, even on a small sample of images. All this requires the search for automatic and without human intervention criteria for assessing similarity.

II. Goal and Motivation

The aim of this paper is to test whether four popular similarity measures (not specially designed for image comparison) could be used to calculate similarity between images. We have tried to do it in our previous paper [9] by comparing the results of similarity measures to the cumulative human perception of similarity between the same images, obtained from an online survey. However, we encountered an enormous problem then – non-uniform judgment and evaluation scales used by the individual respondents.

The survey was designed so that a query image was shown next to a set of sample images, and users were required to specify the exact value of similarity (in their own opinion) in percentage between the query and each image within the set. Since we used nominal, rather than ordinal scale, we have got quite high non-uniformity between individual answers. For example, a respondent specified the similarity between the query and the image X is 80%. Another respondent specified 95% for the same pair of images, while a third respondent specified 40%. Averaging answers having high discrepancies as the above mentioned, could not guarantee reliability and accuracy of obtained "human perception of similarity". So the latter could not be reliably used as a reference.

To test the four similarity measures (Jaccard's index, Euclidean distance, City block distance and Chi-squared dissimilarity) and evaluate how good they are, we decided to use an alternative more objective approach. Inspired by the Top-N accuracy, we applied a similar evaluation approach. We defined a set of 200 images – 50 red roses, 50 tomatoes, 50 red apples and 50 red peppers. The colors of all images are similar – red (roses, fruits, vegetables) and green (leaves). The idea is to check which measure performs best in correct classification of retrieved items (precision) for a specific level of recall. Let's say we are looking for a rose and the system is set up to return 10 results. Then the best similarity measure will be the one that returns most roses out of these 10 results and just a few (or preferably none at all) tomatoes, apples and peppers. However, since there are 50 images of roses, we run 50 queries (every image is used as a query) and average their respective precision for the specified level of recall (top-N results).

This study is important in order to determine which of these four basic similarity measures performs best in searching for images. Results will allow to design and develop an improved universal image retrieval system that could correctly find similar images in various subject domains, or even a system that could automatically select the best similarity measure for a given subject domain by itself.

III. Experimental Environment and Evaluation

The experimental CBIR system used is described in details in [9] and [10]. Briefly, the formation of a feature vector for each image is a sequence of the following actions:

- Each image is divided into 32 by 32 blocks in both width and height dimensions (Fig. 1 to 4).
- All pixels in all the blocks are converted from RGB to one of our 64 primary colors. How these 64 colors were selected and the process of color transformation is described in our previous research [9], [10].
- The dominant color (out of 64 selected colors in our proposed and used color scheme) in each block is determined based on the number of pixels of each color. This dominant color is associated with this block, and the other colors in it are ignored.

In other words, we use just a single color code to substitute multiple pixels per block. In this way, the enormous image color content is reduced to a feature vector with 1024 (32 by 32) color codes. Results of such quantization and color substitution are showed on Fig. 1 to 4. That allows fast image processing and similarity searching. Also, it improves recall as well. The system does the same color analysis for both the image query and the image set and computes such feature vectors for each graphic file. Based on set-theoretic or algebraic methods and similarity measures such as Jaccard Index, Euclidean Distance, City Block Distance and Chi-Square Dissimilarity described in [9] we calculate the similarity factor between the query and each result. At the end, the system returns a sorted list of similar images (Fig. 9).

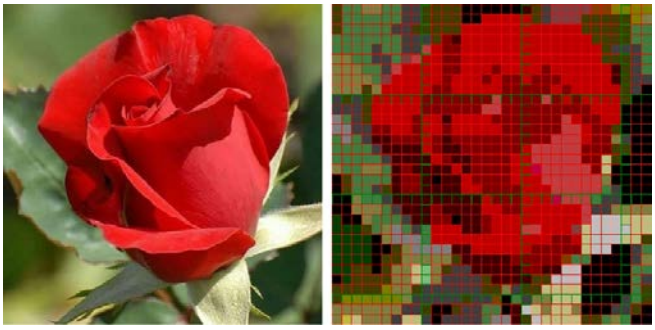


Fig. 1. Original Rose Image Processed by the Application (Left) and its Quantized Image by 32x32 Blocks with Only One Dominant Color Per each Quantization Block (Right).

Fig. 2. Original Tomato Image Processed by the Application (Left) and its Quantized Image by 32x32 Blocks with Only One Dominant Color Per each Quantization Block (Right).

Fig. 3. Original Pepper Image Processed by the Application (Left) and Its Quantized Image by 32x32 Blocks with Only One Dominant Color Per each Quantization Block (Right).

Fig. 4. Original Apple Image Processed by the Application (Left) and its Quantized Image by 32x32 Blocks with Only One Dominant Color Per each Quantization Block (Right).

A set of 200 images (as described earlier) is used in our study. They all have common visual or color characteristics, but are divided in four separate groups - roses, tomatoes, apples and peppers. The feature vectors are stored in the database and each of these 200 images is used as a query in the experiment. It is known which image is from which of the four groups and keeps track of how many images there are in the returned result in the first n similar images from the same group. The first 3, 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 results are examined and the average of the number of images returned from the same group is calculated for every query. This is repeated for each of

the similarity measures with the idea of checking the degree of accuracy/adequacy for each of the measures.

Let's take the group of images of roses for example. It is checked for each rose image, run as a query, how many of the top-N returned results are roses as well. The test is repeated for all 50 rose images and then the average precision value is determined as a proportion of the returned rose images to all returned images (e.g. top-3 results). So, if for "rose image 1" as a query, we have 3/3 returned rose images (i.e. all returned images are roses), for "rose image 2" as a query, we have 2/3 rose images (i.e. 2 images are roses indeed and the third image is another object), and for "rose 3", we have 1/3 rose images returned (i.e. 1 image of a rose and two other objects), then the average precision for top-3 results is (3 + 2 + 1) / (3 + 3 + 3) = 6 / 9 = 2 / 3.

In the experiment, this is done with 50 image queries from each of the four groups and the top-3, -5 ... results are examined. In this way, we can track how the mean precision changes for each similarity measure, based not only on individual queries, but on all 50 queries from each image group. The results are shown in Tables I, (for the set of roses, run as queries), II (for the set of apples), III (for the set of peppers) and IV (for the set of tomatoes).

TABLE I. AVERAGE RESULTS FOR EACH SIMILARITY MEASURE BASED ON ALL 50 ROSE QUERY IMAGES

TOP X RESULTS	Jaccard's index	City block distance	Euclidean distance	Chi-squared dissimilarity
TOP 3 RESULTS	2.6 / 3	2.06 / 3	2.06 / 3	1.82 / 3
TOP 5 RESULTS	4.00 / 5	3.08 / 5	3.16 / 5	2.78 / 5
TOP 10 RESULTS	7.26 / 10	5.50 / 10	5.22 / 10	4.56 / 10
TOP 15 RESULTS	10.08 / 15	7.62 / 15	7.20 / 15	6.24 / 15
TOP 20 RESULTS	12.74 / 20	9.34 / 20	8.82 / 20	7.80 / 20
TOP 25 RESULTS	15.30 / 25	10.98 / 25	10.36 / 25	9.00 / 25
TOP 30 RESULTS	17.96 / 30	12.58 / 30	11.98 / 30	10.36 / 30
TOP 35 RESULTS	20.40 / 35	14.20 / 35	13.76 / 35	11.92 / 35
TOP 40 RESULTS	22.92 / 40	16.06 / 40	15.54 / 40	13.22 / 40
TOP 45 RESULTS	25.06 / 45	17.96 / 45	17.24 / 45	14.64 / 45
TOP 50 RESULTS	27.58 / 50	19.46 / 50	19.14 / 50	16.04 / 50

TABLE II. AVERAGE RESULTS FOR EACH SIMILARITY MEASURE BASED ON ALL 50 APPLE QUERY IMAGES

TOP X RESULTS	Jaccard's index	City block distance	Euclidean distance	Chi-squared dissimilarity
TOP 3 RESULTS	1.76 / 3	1.76 / 3	1.62 / 3	1.60 / 3
TOP 5 RESULTS	2.38 / 5	2.20 / 5	2.04 / 5	1.98 / 5
TOP 10 RESULTS	3.82 / 10	3.42 / 10	3.22 / 10	3.34 / 10
TOP 15 RESULTS	5.24 / 15	4.60 / 15	4.46 / 15	4.58 / 15
TOP 20 RESULTS	6.66 / 20	5.92 / 20	5.46 / 20	5.70 / 20
TOP 25 RESULTS	8.04 / 25	7.08 / 25	6.50 / 25	6.96 / 25
TOP 30 RESULTS	9.20 / 30	8.24 / 30	7.74 / 30	8.18 / 30
TOP 35 RESULTS	10.70 / 35	9.44 / 35	8.72 / 35	9.46 / 35
TOP 40 RESULTS	11.84 / 40	10.68 / 40	9.92 / 40	10.74 / 40
TOP 45 RESULTS	13.06 / 45	11.84 / 45	11.36 / 45	12.08 / 45
TOP 50 RESULTS	14.30 / 50	13.08 / 50	12.36 / 50	13.40 / 50

TABLE III. AVERAGE RESULTS FOR EACH SIMILARITY MEASURE BASED ON ALL 50 PEPPER QUERY IMAGES

TOP X RESULTS	Jaccard's index	City block distance	Euclidean distance	Chi-squared dissimilarity
TOP 3 RESULTS	1.80 / 3	1.38 / 3	1.38 / 3	1.44 / 3
TOP 5 RESULTS	2.32 / 5	1.82 / 5	1.82 / 5	1.70 / 5
TOP 10 RESULTS	4.00 / 10	2.62 / 10	2.54 / 10	2.48 / 10
TOP 15 RESULTS	5.62 / 15	3.70 / 15	3.60 / 15	3.58 / 15
TOP 20 RESULTS	6.94 / 20	4.94 / 20	4.96 / 20	4.80 / 20
TOP 25 RESULTS	8.46 / 25	5.90 / 25	6.14 / 25	5.84 / 25
TOP 30 RESULTS	10.14 / 30	7.00 / 30	7.10 / 30	7.08 / 30
TOP 35 RESULTS	11.42 / 35	8.46 / 35	8.14 / 35	8.44 / 35
TOP 40 RESULTS	12.78 / 40	9.36 / 40	9.08 / 40	9.50 / 40
TOP 45 RESULTS	14.16 / 45	10.54 / 45	10.34 / 45	10.80 / 45
TOP 50 RESULTS	15.46 / 50	11.58 / 50	11.46 / 50	11.96 / 50

TABLE IV. AVERAGE RESULTS FOR EACH SIMILARITY MEASURE BASED ON ALL 50 TOMATO QUERY IMAGES

TOP X RESULTS	Jaccard's index	City block distance	Euclidean distance	Chi-squared dissimilarity
TOP 3 RESULTS	1.92 / 3	2.04 / 3	1.92 / 3	1.86 / 3
TOP 5 RESULTS	2.78 / 5	3.02 / 5	2.78 / 5	2.64 / 5
TOP 10 RESULTS	4.90 / 10	5.12 / 10	4.38 / 10	4.72 / 10
TOP 15 RESULTS	6.62 / 15	7.30 / 15	6.18 / 15	6.44 / 15
TOP 20 RESULTS	8.78 / 20	8.74 / 20	7.74 / 20	8.16 / 20
TOP 25 RESULTS	10.34 / 25	10.32 / 25	9.24 / 25	9.88 / 25
TOP 30 RESULTS	11.82 / 30	11.92 / 30	10.76 / 30	11.68 / 30
TOP 35 RESULTS	13.58 / 35	13.60 / 35	12.36 / 35	13.06 / 35
TOP 40 RESULTS	15.34 / 40	14.90 / 40	13.92 / 40	14.44 / 40
TOP 45 RESULTS	16.66 / 45	16.46 / 45	15.50 / 45	15.78 / 45
TOP 50 RESULTS	18.88 / 50	17.74 / 50	16.78 / 50	16.92 / 50

Results are also presented graphically on Fig. 5 (for the set of rose query images), Fig. 6 (the set of apple query images), Fig. 7 (the set of pepper query images) and Fig. 8 (the set of tomatoes query images).

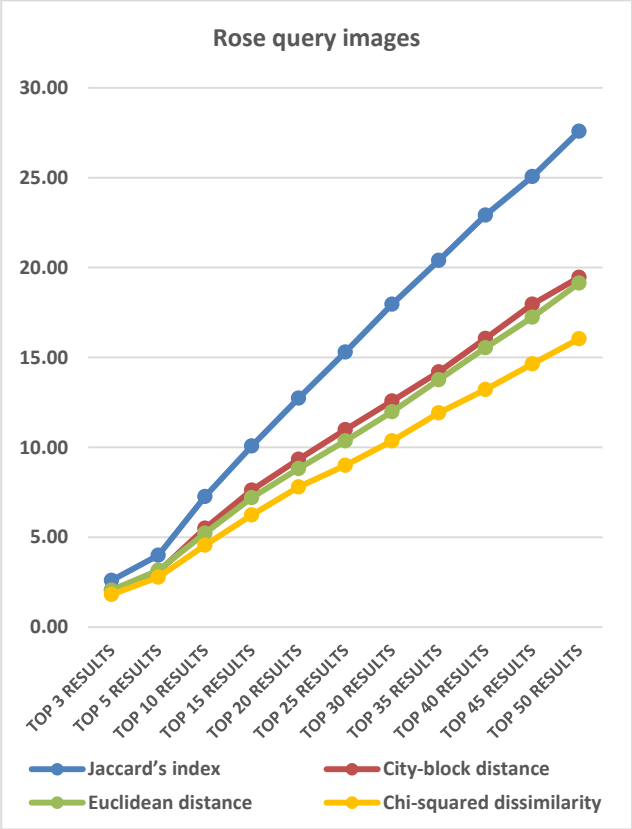


Fig. 5. Average Number of Returned Images, Containing *Roses* (y-axis), Per Similarity Measure and Number of Returned Results (x-axis).

It clearly seems that the Jaccard's index significantly outperforms (retrieves more images containing an object of the same type) all other similarity measures for the set of roses (see Fig. 5). However, this is not the case for tomatoes (Fig. 8), for example, although they have the very same colors. The in-depth analysis of the similarity measures themselves reveals the reason - the Jaccard's index calculates similarity between two images by positionally comparing the dominant colors block by block. All other described measures utilize colors globally. Taking local color distribution into account allows considering not just colors, but shapes and local details as well.

Jaccard's index performs better with roses rather than tomatoes, because the rose's flower consists of multiple individual leafs that reflect light differently and creates dark shadows between leafs (Fig. 1), while tomatoes are singular convex rounded objects (Fig. 2). So, accounting position of the shadows, the Jaccard's index can more easily and reliably guess if the red object in the center of the image is a rose or something else. However, distinguishing a convex red tomato from a convex red apple is much more difficult. That is why Jaccard's index outperforms all other similarity measures for the set of roses (due to additional surface features – shadows between leafs) and the set of peppers (due to the oblong shape), but achieves less better (but still better) results for apples, and no improvement for tomatoes image set.

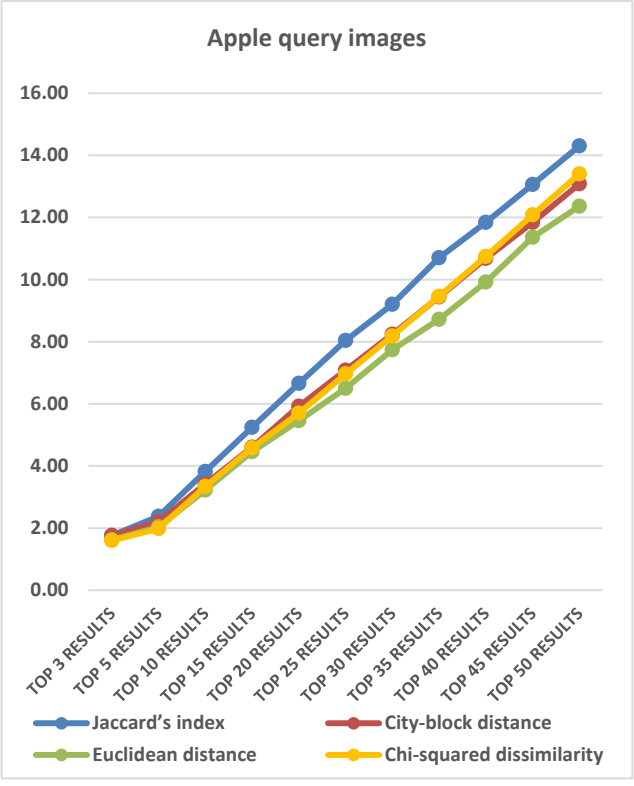


Fig. 6. Average Number of Returned Images, Containing *Apples* (y-axis), Per Similarity Measure and Number of Returned Results (x-axis).

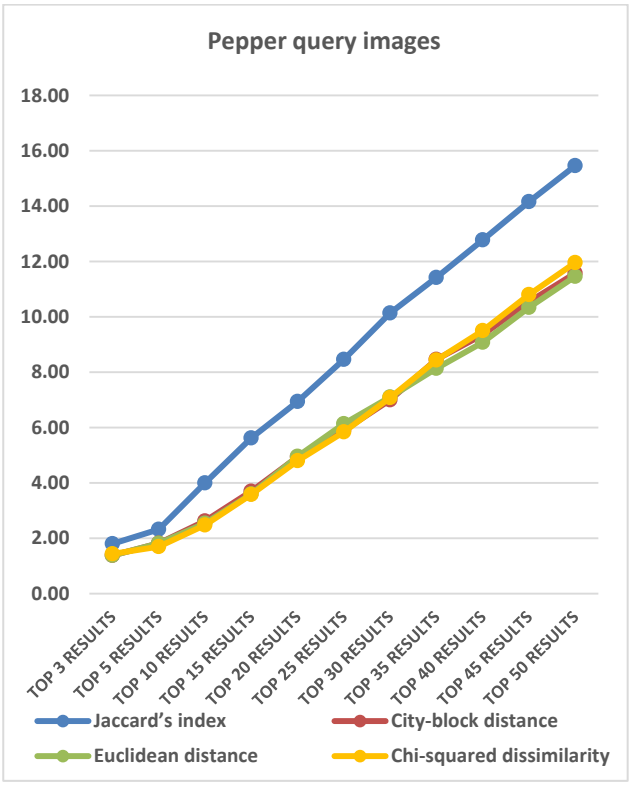


Fig. 7. Average Number of Returned Images, Containing *Peppers* (y-axis), Per Similarity Measure and Number of Returned Results (x-axis).

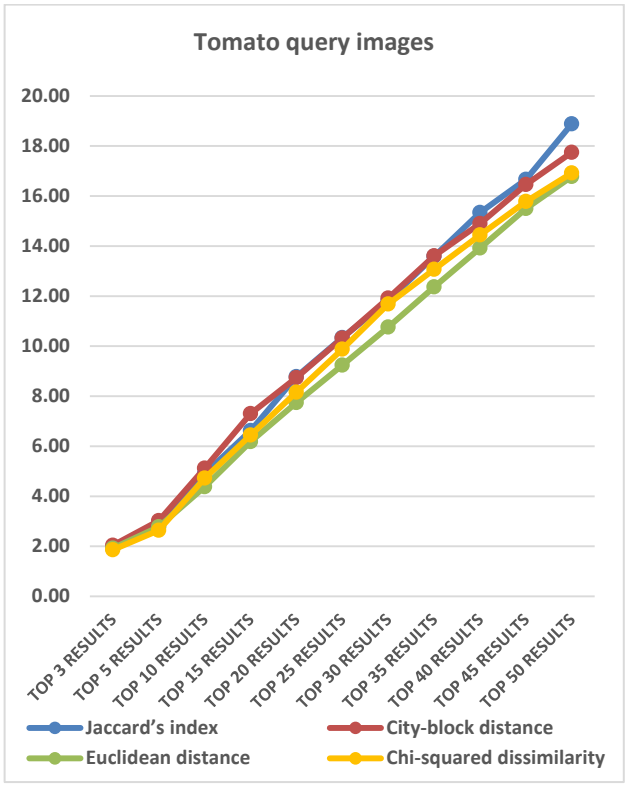


Fig. 8. Average Number of Returned Images, Containing *Tomatoes* (y-axis), Per Similarity Measure and Number of Returned Results (x-axis).

Fig. 9. Example of Top-50 Similarity Results based on Rose Image Query and Jaccard's Index Sorted by Degree of Similarity in Descending Order.

IV. Conclusion

Results from the series of experiments show that:

- All described similarity measures (Jaccard's index, Euclidean distance, City block distance and Chi-squared dissimilarity) could be used to calculate similarity between images. They all provide similarity within the range of [0, 1] and allow search results to be sorted by similarity in reverse order.
- When searching by color content, and consider colors globally, then Euclidean distance, City block distance and Chi-squared dissimilarity produce commensurate results. That is clearly noticeable on Fig. 5 to 9. The main difference between these metrics is in the magnitude of the calculated value. However, the relationships between the calculated similarity factors remain the same, regardless of which one of these three similarity measure is used. It should be noted here that exactly the relationships between similarity factors, rather than the absolute values themselves, create the order of the search results.
- In contrast to all other similarity measures, Jaccard's index compare feature vectors positionally, so it takes into account not just colors, but also their spatial distribution. As a result, it indirectly considers shape, position, orientation and other features.
- When objects have specific features on their surfaces or irregular (e.g. oblong) shape, then Jaccard's index significantly outperforms other similarity measures. That is easily noticeable on Fig. 5 for the set of roses and Fig. 7 for the set of peppers.
- In general, when there is no a-priori information about the image database, the Jaccard's index seems the best single similarity measure between images. This statement is supported by the data in all tables and figures.

Acknowledgment

This paper is supported by project 2022–EEA–01 "Analysis of big data processing algorithms and their application in multiple subject domains", funded by the Research Fund of the "Angel Kanchev" University of Ruse.

References

[1] Liu, Y., Zhang, D., Lu, G., & Ma, W. Y. (2007). A survey of content-based image retrieval with high-level semantics. Pattern recognition, 40(1), 262-282.

[2] Long, F., Zhang, H., & Feng, D. D. (2003). Fundamentals of content-based image retrieval. In Multimedia information retrieval and management (pp. 1-26). Springer, Berlin, Heidelberg.

[3] Shivamurthy R C, Procedures Design and Development of Framework for Content Based Image Retrieval, International Journal of Advanced Research in Engineering and Technology, 12(1), 2021, pp. 1167-1180.

[4] Chugh, H., Gupta, S., Garg, M., Gupta, D., Juneja, S., Turabieh, H., ... & Kiros Bitsue, Z. (2022). Image retrieval using different distance methods and color difference histogram descriptor for human healthcare. Journal of Healthcare Engineering, 2022.

[5] Ashraf, R., Ahmed, M., Jabbar, S., Khalid, S., Ahmad, A., Din, S., & Jeon, G. (2018). Content based image retrieval by using color descriptor and discrete wavelet transform. Journal of medical systems, 42(3), 1-12.

[6] Xu, G., Xiao, K., & Li, C. (2019). Shape description and retrieval using included-angular ternary pattern. Journal of Information Processing Systems, 15(4), 737-747.

[7] Bu, H. H., Kim, N. C., Park, K. W., & Kim, S. H. (2019). Content-based image retrieval using combined texture and color features based on multi-resolution multi-direction filtering and color autocorrelogram. Journal of Ambient Intelligence and Humanized Computing, 1-9.

[8] Mistry, Y., Ingole, D. T., & Ingole, M. D. (2018). Content based image retrieval using hybrid features and various distance metric. Journal of Electrical Systems and Information Technology, 5(3), 874-888.

[9] Marinov M., I. Valova, Y. Kalmukov, "Comparative Analysis of Existing Similarity Measures used for Content-based Image Retrieval", 2019 X National Conference with International Participation (ELECTRONICA), Sofia, Bulgaria, 16 - 17 May 2019.

[10] Marinov, M., Valova, I., & Kalmukov, Y. (2020). Design and implementation of the CBIR system for academic/educational purposes. In 2020 International Conference Automatics and Informatics (ICAI) (pp. 1-4). IEEE.

Automatic Detection of Roads using Aerial Photographs and Calculation of the Optimal Overflight Route for a Fixed-wing Drone

Miguel Pérez P, Holman Montiel A, Fredy Martínez S
Facultad Tecnológica, Universidad Distrital Francisco José de Caldas
Bogotá, Colombia

***Abstract*—Currently, fixed-wing drones have become indispensable tools for the surveillance of large areas of land, justified by their better cost/benefit ratio, great flight autonomy, and payload capacity. In particular, the identification of roads, traffic control, monitoring of wear on asphalt layers, risk identification, and safety improvement are applications that are being implemented in these unmanned aerial vehicles. Tracking a road requires systems capable of detecting artificial marks through images employing aerial photographs that allow the implementation of optimal overflight routes. This research work presents a solution to the problem of road tracking from aerial photographs and implements an image processing algorithm and morphological techniques that calculate and traces the ideal route for the drone to track automatically, regardless of its orientation and the type of road.**

***Keywords*—Automatic road tracking; decorrelation stretching; aerial imagery; optimal overflight route calculation; UAV**

I. Introduction

Transportation infrastructure, and in particular roads, are of vital importance for the growth and development of a country, and even more so for a country like Colombia, since 80% of the country's cargo is moved by road [1]. Surveillance of these important roads is normally done through simple methods such as cameras or with the public forces located in different segments of these roads. These methods are costly due to the many elements or people that must be available [2], [3]. Currently, the need to implement new, less expensive methods for surveillance has forced researchers to investigate systems that use fixed-wing drones for this task [4]. Such systems take advantage of the great autonomy and high payload capacity of these drones [5] for the process of monitoring and surveillance of roads. This monitoring process cannot be teleoperated due to the extension of the territories to be monitored; for this reason, research is aimed at autonomous systems that use digital image processing algorithms, obtained by the Unmanned Aerial Vehicles (UAV) themselves; to automatically detect the road and thus be able to extract elements of interest autonomously [6], [7].

Asphalt roads can be considered extraneous or artificial features that extend for many kilometers in a natural landscape [8], [9]. A process used to highlight these landmarks in an image is called decorrelation stretching (DCS) [10], [11], [12]. The DCS technique is complemented with color detection [13], [14] and image filtering [15] algorithms; which extract the violet areas corresponding to the road, leaving it in a binarized image. It should be clarified that the images used in this research are aerial photographs of roads taken by fixed-wing drone overflight; where there are great differences between them (altitude taken from the photo, direction, light conditions and brightness, topography, landscape, etc.) which demonstrates the robustness and efficiency of the developed system. Once the road is detected, the system can calculate the extreme points [16] and trace the direction of this obtaining the optimal route of overflight of the UAV [17] that allows following the road for many km.

To recapitulate, the proposed system incorporates a decorrelation stretching algorithm and filtering and color detection algorithms to automatically detect a road in an aerial image. In addition, it proposes an algorithm that allows computing the optimal overflight path of the detected road by implementing the endpoints of the detected road. These contributions are described in detail in the following sections, which are organized as follows: Sections II and III present the general definitions and methodology used respectively, and section IV presents the results obtained.

II. Methodology

This research contributes to a joint operation of decorrelation stretching algorithms (DCS), image filtering, and color detection algorithms together with image processing libraries developed entirely in MATLAB. They allow detecting and calculating a flyover route for roads using aerial images. A description of each of the elements used is presented below.

A. Decorrelation Stretching (DCS)

Decorrelation stretching is a process that improves the color differences found in an image by removing the inter-channel correlation found in the image pixels. Decorrelation stretched images provide an overview that improves spectral reflectance variations; that is, it improves color differences in images with high inter-channel correlation [11], [12].

Therefore, using DCS is useful for intensifying the color difference between roads and the surrounding landscape in an image, which facilitates the detection of roads. Fig. 1 shows an image of a road subjected to the decorrelation stretching process implemented in this research; in it, you can see the color difference between Fig. 1 (a) and (b), in addition, you can see the violet hue that takes the road to be detected.

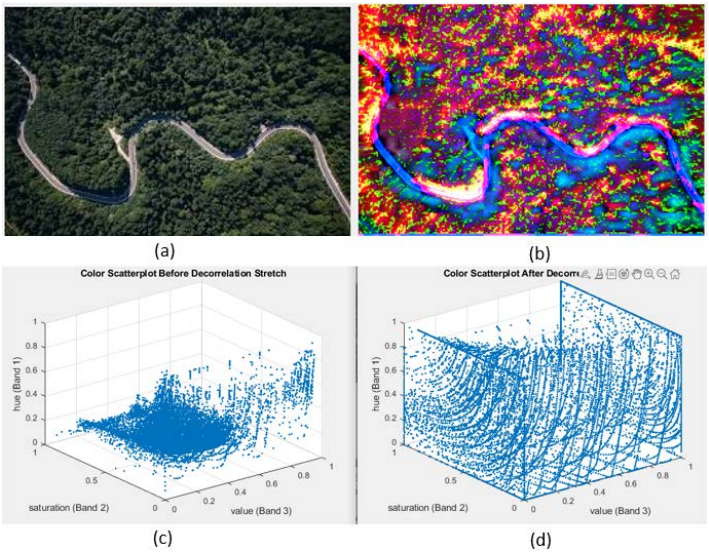


Fig. 1. (a) Aerial View of the Road from Stock Photo 2024801798 [18], (b) Image Processed by the Decorrelation Stretching Algorithm, (c) Point Spread Cloud of the Original Image, (d) Point Spread Cloud of the Resulting Image.

The decorrelation stretching is a linear operation in pixels in which the parameters depend on the values of the original image statistics and the target image and can be expressed using the following equations [19].

$$b = T * (a - m) + m\_target \quad (1)$$

a and b are vectors of n bands times 1, T is the linear transformation matrix and m contains the mean of each band in the image, m_target are vectors of n bands times 1 containing the desired output mean in each band. The linear transformation matrix T is calculated by performing a proper decomposition of the correlation matrix as follows:

$$Corr = inv(SIGMA) * Cov * inv(SIGMA) \quad (2)$$

$$S(k,k) = \frac{1}{\sqrt{LAMBDA(k,k)}} \quad (3)$$

Where $S(k,k)$ is a diagonal matrix containing the stretching factors for each band; finally, the matrix T used in (1), is calculated through:

$$T = SIGMA\_target\ V\ S\ V'\ inv(SIGMA) \quad (4)$$

Although the decorrelation stretching process is usually implemented in RGB format images [20]; the operations proposed here can be applied to any multiplane matrix. Therefore, the algorithm can be applied to images in other formats. For this research the image was processed in HSV format; since it takes advantage of the high stability of this format to the changes of brightness and luminosity in the image; which benefits the detection [21].

In Fig. 1 (c) and (d), you can see the cloud of points representing the dispersion of pixel values in both the original and the resulting image; image (c) is much less dispersed than (d); which demonstrates the color differences between the two images, eliminating the high correlation between channels.

B. Color Detection and Image Filtering Algorithm

Color detection is a technique used to separate objects containing colors of interest into the different existing color spaces; it is important to clarify that color spaces are nothing more than a representation of how machines understand the existing color in the real world [22]. An image in RGB format is made up of three matrices representing the level of red, green, and blue in the image; therefore, the combinations of these three matrices give rise to a wide range of colors ranging from black to white. A color segmentation algorithm reads the three matrices that make up the image and extracts from them the pixels containing the target color; usually, the output of this algorithm is a single plane or binarized matrix which contains in white the pixels corresponding to the target color and in black the other pixels; which are considered as the background of the image (see Algorithm1).

Algorithm 1. Color Detection

```
Start Variables;
create an array of zeros the size of the image
For a=1: up to the number of rows
For b=1: up to the number of columns
If (Image (a,b,R)> threshold)
If (Image (a,b,B)> threshold)
If (Image (a,b,G)< threshold)
transform into 1 pixel with the target color
End If
End If
End If
End For
End For
```

Algorithm 1 shows the implemented color detection; in this algorithm, it is observed that the function goes through the color image and transforms the pixels corresponding to the target color. This procedure is achieved through a function that defines the transformation threshold, see Fig. 2.

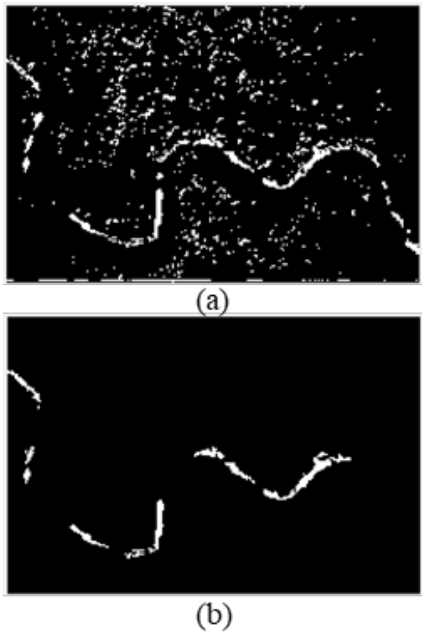


Fig. 2. (a) Binarized Image Employing the Color Detection Algorithm, (b) Filtered Image.

From Fig. 2(a) the need for the image filtering algorithm is evident. The process of image filtering is that which removes unwanted objects or noise from the original image. There are innumerable image filtering methods or functions [23]; but one of the most widely used is bwareaopen (BinaryImage, minPixels) object size filtering. This function removes objects from a binarized image that contain less than the minimum number of pixels specified by minPixels; such a filtering process finds the connected components of the binarized image and calculates the area (in pixels) for further removal. Another filter implemented is the imfill function (BW, locations), which performs a flood fill operation on the background pixels, and is used to eliminate noise in the detected objects. It should be noted that this function removes all black objects surrounded by white elements in the image. Fig. 2 (b) shows the result of the image filtering algorithms implemented in this research.

III. Implementation

The developed system is composed of two main operational blocks: the first block oversees detecting and extracting the existing roads in the processed image; this block makes use of the technique of stretching by decorrelation and conventional image filtering, see Fig. 3. The second block aims to develop an algorithm that calculates the optimal overflight route for a UAV to follow, this algorithm will be explained in full in the next section and is a creation entirely developed by the authors of this research.

A. Overflight Route Calculation for Autonomous Navigation

Calculating the overflight route that allows autonomous navigation for extended periods is essential in video surveillance systems [24] [25]; this calculation must draw a straight line between the start and end points to maximize the capabilities of the fixed-wing drone and minimize its fuel consumption. The problem is that most roads are not straight (see Fig. 4 (a)), therefore the system must calculate the orientation (horizontal, vertical, diagonal, etc.) and direction of the road to calculate an overflight route (in a straight line) that allows overflying the entire road.

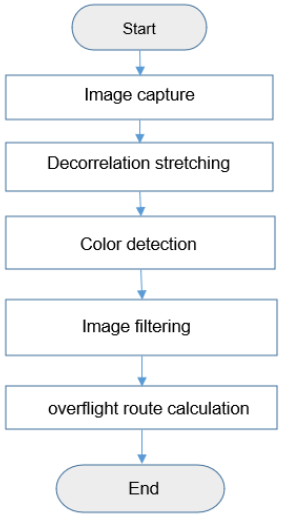


Fig. 3. Block Diagram of the Developed System.

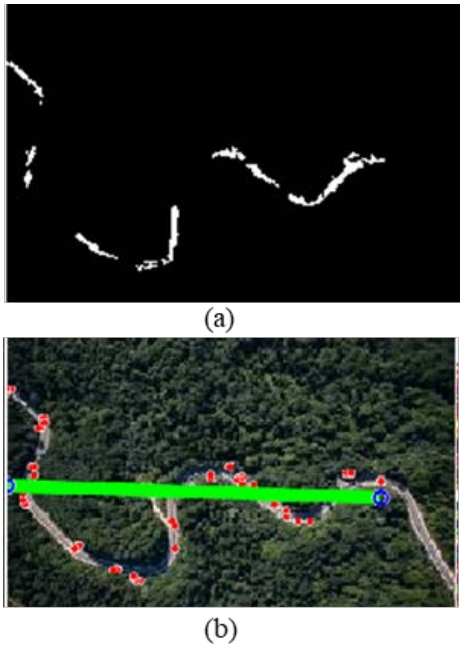


Fig. 4. (a) Road Detected by the Algorithm, (b) Automatic AUV Overflight Route Calculation.

Fig. 4(a) shows the result of the road detection algorithm. Here it can be seen that the road is not detected as a single element; but as multiple elements that are added together to give the totality of the road. These interruptions in the continuity are due to external elements that prevent its detection and with which the route calculation algorithm must deal. The proposed overflight path calculation algorithm starts by identifying the endpoints of each element, which can be seen in red in Fig. 4(b). The function in charge of calculating the endpoints by morphological processing is called

regionprops (ImageBina,'Extreme') [26]. Then the image is divided into four equal quadrants, and it is calculated which of these extreme points for each quadrant is higher, lower, left, and right than the others; which determines the orientation and direction of the road. Once these points have been calculated, the start and end point of the route is calculated as the midpoint between these points, see Algorithm 2.

Algorithm 2. Navigation route calculation

```
Start Variables;
for a=1: up to the number of items detected
calculations of the extreme points of each element
 if (the element is the highest)
                save highest_item
 else if (is the lowest)
                save lowest_item
 else if (is the most to the left)
                save left_item
 else if (it is the most to the right)
                save right_item
 end if
end for
%%% horizontal or vertical direction calculation

if (highest_item < left_item)
 fI=fminI+(fmaxI-fminI)/2  % top point
else
 fD=fminD+(fmaxD-fminD)/2 % left point
end if
if (lowest_item > right_item)
 cU=cminU+(cmaxU-cminU)/2 % down point
else
 cD=cminD+(cmaxD-cminD)/2 % right point
end if
paint line on the detected points
```

IV. Results

The algorithm was implemented on a computer with Windows 10 Pro 64bits operating system, Intel(R) Core (TM) i3-3110M CPU 2.40GHz (4 CPUs), RAM 12288MB, and MATLAB 2020b Update 5 under license from Universidad Distrital Francisco José de Caldas. The implemented system operates pixel by pixel on the input matrix. Table I shows the time report; here it can be seen that the time that contributes most to the system depends on the number of elements that were detected; since the algorithm for calculating extreme points and calculating the route must be applied individually to these elements.

The automatic road detection and overflight route calculation system was tested with different images taken by fixed-wing drone overflight. These images were taken with different cameras at different altitudes and different times of the day to demonstrate the robustness of the algorithm.

Fig. 5 shows the response of the developed system; it is observed that the system can detect and calculate the overflight path for roads going vertically, horizontally, or diagonally; demonstrates the effectiveness of the algorithm for images with different illuminations; double-lane roads, and cars. Finally, demonstrate what happens to the algorithm if it encounters intersections, elevated bridges, round-point, and other elements that are usually present on the road.

As the algorithm is designed to run on a mini pc on board the aircraft, the validation of operation was performed using different processing units. Fig. 6 shows the response time of the algorithm on different processors; it should be noted that these times were measured using the same image in which five elements were obtained after binarization.

TABLE I. System Processing Time Measurements

Time Report [seconds]		
Image size in pixels	**Number of Detected Objects**	**Time in Seconds**
33.000	3	0.2488
50.246	1	0.1113
50.246	2	0.3273
50.325	3	0.2526
50.400	1	0.0977
50.400	3	0.4041

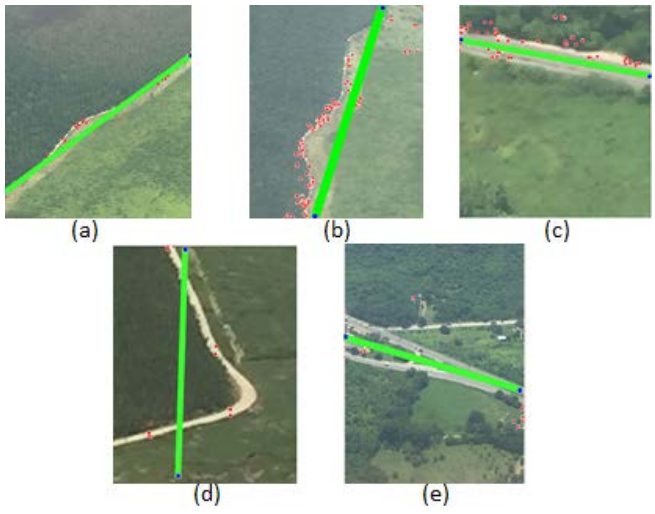


Fig. 5. Tests of the Algorithm with Photos taken from Different Environments and Altitudes.

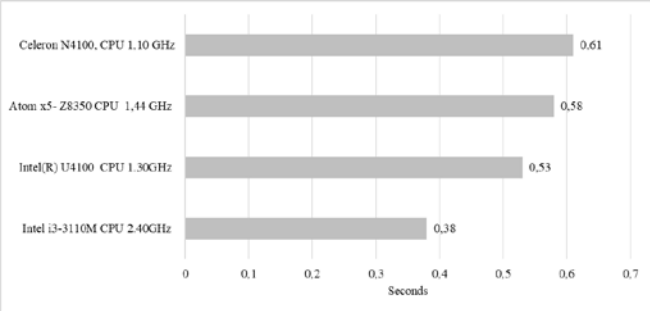


Fig. 6. Algorithm Response Time with Different Processors.

V. Conclusions and Future Work

As can be seen in Fig. 5, the performance of the system is high since it can detect the road in photos taken with different devices at different heights and with different degrees of illumination. It is important to emphasize that the overflight path calculation algorithm is a novel proposal of the researchers, this calculation is performed through the endpoints of the detected road.

Although the decorrelation stretching algorithm has been implemented to highlight the color difference in an image; its use to detect roads in aerial images is a proposal introduced in this research. This proposal demonstrates its robustness since it can separate the colors of the road and the landscape in different aerial images.

Finally, as can be seen in Fig. 5(e) the system can make a coherent decision when encountering intersections, elevated bridges, round-point, and other elements that are usually present on the road. Thus, demonstrating the high degree of stability and applicability of the proposed solution; ensuring that the road detection system can be applied in various environments since it can be functionally adapted to various road environments.

Although the algorithm was tested in a specific task, this implementation offers the possibility that it can be used in various fields of the productive sector. Since the proposed solution allows for the identification clearly and punctually of artificial long objects in a natural landscape. The next step is to complement the system with algorithms that can automatically detect anomalies (pipeline monitoring, fire outbreaks, traffic jams, suspicious vehicles, etc.) that may occur during the overflight of the drone, and thus seek to complete an autonomous aerial tracking and surveillance system.

Acknowledgment

This work was supported by the Universidad Distrital Francisco José de Caldas, in part through CIDC, and partly by the Technological Faculty. The views expressed in this paper are not necessarily endorsed by the university. The authors thank the research group ARMOS for supporting the development of the code and its implementation.

References

[1] A. M. Orozco, " Informe Transporte en cifras Estadisticas 2020", in Ministerio de transporte de colombia, centro de estudios regionales, Versión 1.0, 2021.

[2] M. Al-Smadi, K. Abdulrahim, R. Abdul, "Traffic Surveillance: A Review of Vision Based Vehicle Detection, Recognition and Tracking", *International Journal of Applied Engineering Research*, vol. 11, no. 1, pp 713-726, 2016.

[3] T.M. Amir-Ul-Haque, M. Das, S. Reza, "Computer Vision Based Traffic Monitoring and Analyzing From On-Road Videos", *Global Journal of Computer Science and Technology*, vol. 19, no. 2, pp. 1-7, 2019.

[4] P. Vanitchatchavan, "Low-Cost Smart Surveillance and Reconnaissance Using VTOL Fixed Wing UAV", *IEEE Aerospace Conference*, 2020.

[5] S. A. Hassnain Mohsan, M. Asghar Khan, F. Noor, I Ullah, M. H. Alsharif, "Towards the Unmanned Aerial Vehicles (UAVs): A Comprehensive ReviewGait", *Drones*, vol. 6, no. 147, pp. 1-27, 2022.

[6] A Berrondo Urruzola, "Detección de carreteras en imágenes de reconocimiento remoto mediante deep", Grado en Ingeniería Informática Computación, Univeridad del pais vasco, Facultad de informatica, 2020.

[7] A. Yasin, A. Kocatepe, "Automatic road detection from orthophoto images", *mersin photogrametri journal*, vol. 2, no. 1, pp. 10-17, 2020.

[8] E. Brewer , J. Lin, P. Kemper, J. Hennin, D. Runfola, "Predicting road quality using high-resolution satellite imagery: A transfer learning approach", *PLoS ONE*, vol. 16, no.7, pp. 1-18, 2021.

[9] E. Garilli,N. Bruno, F Autelitano, R Roncella, F Giuliani, "Automatic detection of stone pavement's pattern based on UAV photogrammetry", *Automation in Construction, vol 122*, pp. 1-14, 2021.

[10] R. Nanmaran, S. Hari Priya, "Design and Development of Decorrelation Stretch Technique for Enhancing the Quality of Satellite Images with Improved MSE and UIQI in Comparison with Wiener Filter", *ECS Transactions*, vol. 107, no. 1, 2022.

[11] P. Sinha, B. Horgan, R. Ewing, E. Rampe, M. Lapotre, M. Nachon, M. Thorpe, A. Rudolph, C. Bedford, K. Maso, E. Champion, P. Gray, E. Reid, M. Faragalli, "Decorrelation stretches(DCS) of visible images as a tool for sedimentary provenance investigation on earth and mars", *NTRS - NASA Technical Reports Server*, March 16, 2020.

[12] S. Boopathiraja, P. Kalavathi, M. Geethalakshmi, "Performance Analysis of Multispectral Color Composite Image Enhancement Technique using Decorrelation Stretching and Histogram Equalization Methods," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 4, pp.319-323, 2018.

[13] P. Sudharshan & M. Deepa, "Color detection in RGB-modeled images using MATLAB", International Journal of Engineering and Technology, vol. 7, no. 2, pp. 29-33, 2018.

[14] A. Hasan Ali, M. Rasheed, S. Shihab, T. Rashid, A. AbdulJabbar, S. Hussein Abed Hamad, "An Effective Color Image Detecting Method for Colorful and Physical Images", *Journal of Al Qadisiyah for Computer Science and Mathematics*, vol. 13, no. 1, pp 88 – 98, 2021.

[15] A. Raghunandan, Mohana, P. Raghav and H. V. R. Aradhya, "Object Detection Algorithms for Video Surveillance Applications", *International Conference on Communication and Signal Processing (ICCSP)*, pp. 0563-0568, 2018.

[16] R. Parekh, "Fundamentals of image, audio, and video processing using matlab® with applications to pattern recognition", Edition1st Edition, First Published202, eBook Published15 April 2021.

[17] Y. Chen, X zhou, "research and implementation of robot path planning based on computer image recognition technology", *Journal of Physics: Conference Series*, 1744 022097, 2021.

[18] Shutterstock. Image repository. https://www.shutterstock.com/es/image-photo/aerial-view-on-mountain-road-drone-2024801798. 2022.

[19] A. Kumar Shakya, A. Ramola, A. Vidyarthi, K. Sawant, "Satellite Image Enhancement for Small Particle Observation using Decorrelation Stretcher", *International Conference on Advances in Computing, Communication & Materials (ICACCM)*, 2020.

[20] The MathWorks Inc, "Image Processing Toolbox For Use with MATLAB®", decorstretch function, Version 3, User's Guide, https://www.mathworks.com/help/images/ref/decorrstretch.html.

[21] D. Hema, S. Kannan. "Interactive Color Image Segmentation using HSV Color Space", *Science and Technology Journal*, vol. 7, no. 1, pp. 37-41, 2020.

[22] V. Tiwari, V. Sharma, Development of Algorithm for Object Detection & Tracking Using RGB Model International Journal of Computer Trends and Technology ·, vol. 8, no. 2, pp. 37-44, 2020.

[23] B. Desai, U. Kushwaha, S. Jha," Image Filtering- Techniques, Algorithm and Applications", *GIS Science Journal*, vol. 7, no. 11, pp. 970-975, 2020.

[24] A. R. Kuroswiski, N. M. F. de Oliveira and É. H. Shiguemori, "Autonomous long-range navigation in GNSS-denied environment with low-cost UAV platform", Annual IEEE International Systems Conference (SysCon), 2018.

[25] S. Mufti, V. Roberge, and M. Tarbouchi, "A GPU Accelerated Path Planner for Multiple Unmanned Aerial Vehicles", *IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, 2019.

[26] J. Ananthanarasimhan, P. Leelesh, M .S. Anand, R. Lakshminarayana, "Validation of projected length of the rotating gliding arc plasma using 'regionprops' function", *Plasma Res. Express*, vol. 2, no. 3, 2020.

Research on Regional Differentiation Allocation Mode of Energy Finance based on Attention Mechanism and Support Vector Machine

Ling Sun[1]*
School of Economics and Management
Nanjing Vocational University of Industry Technology
Nanjing Jiangsu 210023, China

Hao Wu[2]
Jiangsu Suning Bank
Nanjing Jiangsu 210019, China

***Abstract*—This paper studies the prediction method of regional differentiated allocation mode of energy finance based on attention mechanism and support vector machine to provide scientific guidance for the future development direction of energy finance in each region. Analysis of the key factors influencing the energy consumption, through the attention mechanism to extract the regional factors such features constitute the details of the sample set, the characteristics of the sample set after implementation of fusion and normalized processing, gain new characteristics of sample set as input to construct support vector machine forecasting model, prediction of energy consumption in each region of the output. According to the results, the differentiated allocation patterns of energy finance in each region are predicted. The results show that the prediction model of this method has high training and test prediction accuracy, and the prediction results are consistent with the actual data in historical statistics. Compared with the existing methods, the method of this study can more scientifically and effectively predict the sustainable and stable development of energy finance in various regions of the city in the future. The energy consumption of the experimental city predicted in this study in the next nine years is from high to low in the order of region C, region a and region B. from this, it is predicted that the regions A, B and C of this city in the future will be applicable to the government market dual oriented Government oriented and market-oriented energy finance allocation models. The prediction results can provide scientific guidance for the sustainable and stable development of energy finance in various regions of the city in the future.**

Keywords*—Attention mechanism; support vector machine; energy finance; differentiation configuration mode; energy consumption

I. INTRODUCTION

Energy finance refers to a cooperative model based on the energy system that uses the financial system as a driving force after the integration of energy and finance. It is a brand-new model that promotes the coordinated development of energy and finance [1]. Energy finance is mainly based on the energy industry chain. With the help of financial means, it studies the interaction between the new energy industry and the financial industry from three aspects: the initial financing, the intermediate integration of resources, and the final realization of value-added, and puts forward corresponding safeguard measures. The energy industry is usually dependent on the financial industry in its development. To obtain the highest profit and achieve the highest capital utilization rate, the financial industry also needs to rely on the assistance of the energy industry [2,3]. As the main body of energy consumption in China, the consumption of primary energy in my country has shown an increase year by year in recent years. On the premise of maintaining the sustained and stable development of our country's energy finance, the development level of energy finance is gradually improved in various regions, so that can achieve the coordinated development of energy and finance in various regions. It is necessary to configure a differentiated energy financial model that meets its development characteristics for each region of our country [4]. Therefore, it is necessary to effectively combine the respective advantages of energy and finance in each region to effectively alleviate the financial problems faced by energy development. Meanwhile, new energy financial development paths will be provided to all regions. Regarding the question of how to configure a differentiated energy finance model that applies to each region, the most important thing is to accurately predict the energy consumption of each region. Based on the prediction results, the energy financial allocation models applicable to different regions can be analyzed [5].

The attention mechanism belongs to a type of model that simulates the attention mechanism of the human brain. Its principle is to use the method of attention probability distribution calculation to highlight the effect of a certain main input on the output, and to obtain more feature detail information. When the attention mechanism is used in deep learning models, it can help improve the overall accuracy and efficiency of such models [6, 7]. Meanwhile, the combination of attention mechanisms and machine learning algorithms to solve problems in projects or life has become a development trend. Support vector machine is a relatively important prediction algorithm in machine learning. It can transform the input in the low-dimensional space into the high-dimensional space through the kernel function and slack variables, and obtain the linear sample data in the best classification method [8]. Support vector machine has the advantages of being able to find the optimal solution globally, strong learning ability, and strong pan-China ability. It can be widely used in multi-dimensional function prediction and various recognition problems [9].

*Corresponding Author.

The attention mechanism and machine learning methods are applied to study the regional differentiation of energy finance, which can realize the prediction of the regional energy financial differentiation allocation mode. However, the research should analyze the key influencing factors that affect energy consumption and extract the relevant key influencing factors on the basis of a full analysis of the regionally differentiated configuration mode of energy finance. The data under the key influencing factors are formed into a sample set, and the data of the sample set is feature-fused using the normalization method, so that can realize the establishment of the support vector machine feature model. In the entire process, key factors can be determined in terms of economic development degree, urbanization degree, energy efficiency application degree, industrial structure and demographic factors that affect energy consumption, which can realize scientific and effective forecasts. Traditional energy finance regional differentiation research is mainly to analyze the correlation between energy industry projects and energy financial derivatives in various aspects. The process of using correlation to achieve prediction requires constant adjustment of model parameters to improve the accuracy of prediction, and the process is more complicated. This paper combines the attention mechanism and support vector machine to build a scientific prediction model, which can simplify the prediction process. Through the identified key factors, the sample set is input into the support vector machine prediction model, the energy consumption prediction results of different regions are output, and the results of the model output are used to analyze the energy financial configuration modes applicable to different regions. Through the identified key factors, the sample set is input into the support vector machine prediction model, the energy consumption prediction results of different regions are output, and the results of the model output are used to analyze the energy financial configuration modes applicable to different regions.

Therefore, the purpose of this paper is to combine the attention mechanism and support vector machine to achieve a scientific and accurate prediction of the regional energy finance differential distribution mode, and can guide the development direction of Regional Energy Finance in the future according to the prediction results, promote the sustainable and stable development of Regional Energy Finance in the future, and make up for the shortcomings of existing research.

II. Related Work

The concept of energy finance appeared in the 1880s. As the traditional energy structure upgrades and the capital demand for corporate development has grown, energy finance has gradually developed. However, the issue of energy finance has become the core issue of today's world economic development. It is not only closely related to the security of the country, but also closely related to the development of society [10]. Therefore, the differential analysis of energy finance between countries or regions has gradually become one of the important research hotspots. Many relevant experts and scholars have conducted analysis and research on the differences in energy finance. Some researchers have studied the framework of regional energy finance cooperation under the new normal [11], and they theoretically explain the importance of forecasting functions to regional energy finance. Some scholars use the GABP algorithm to establish an energy financial risk early-warning model to realize the overall risk prediction of energy finance, but the accuracy of the prediction still needs to be considered [12]. Some researchers analyze the necessity of the association between early warning models and regional energy finance from a theoretical perspective [13]. In terms of the research on the regionally differentiated allocation model of energy finance, most of the existing relevant literature is theoretically analyzed. Some researchers use provincial panel data to analyze financial risks from a differentiated perspective and analyze the potential impact of regional financial risks [14]. Some researchers use the VAR model to study the regional differentiation of rural finance and analyze the correlation between influencing factors [15]. However, the current research on the difference of energy finance is still one-sided, and there are few methods to output scientific and accurate prediction results from the data accumulated in history.

The application of machine learning methods in the financial field is more common, it mainly has stock forecasting, quantitative finance, investment portfolio analysis, etc. [16]. However, there are few studies on regional financial differentiation, which mainly focus on the collection of regional financial data and regional financial risk index. A fuzzy comprehensive evaluation or analytic hierarchy process is applied to evaluate regional financial risks. The analysis of the difference is mainly based on the risk ranking by the evaluation grade [17-18]. There are also a few researchers who study the stability and differences of regional finance [19-21]. However, this type of research is also limited to the use of evaluation methods to obtain results. Few machine learning methods are applied to difference analysis, and data-driven methods are not used to output prediction results.

In summary, the machine learning method has been applied in the financial field, but the existing research on regional financial differentiation mainly focuses on the theoretical level and the correlation analysis between factors. Some studies have made key determination on the economic development degree, urbanization degree, energy efficiency application degree, industrial structure and population factors that affect energy consumption; however, the prediction of regional differentiation of energy finance by means of machine learning is still in the development stage.

Therefore, on the basis of the existing research results and after analyzing the key factors affecting energy consumption, this study uses the attention mechanism to extract the detailed features of these factors in each region, form a feature sample set, and fuse and normalize the feature sample set to obtain a brand-new feature sample set input, build a support vector machine prediction model, and realize scientific and effective research and Analysis on the basis of the existing research, provide auxiliary decision support for relevant management departments.

III. Analysis of the Differential Configuration Mode for Energy Finance Region

To achieve the coordinated development of energy and finance in various regions, it is necessary to configure

differentiated energy financial models in line with their development characteristics for different regions to achieve a more effective combination of the respective advantages of energy and finance. Its model can not only effectively alleviate the financial problems faced in energy development, but also provide brand new development paths to the financial industry in various regions [22]. Generally, the regional energy financial allocation mode can be divided into three types: government-oriented regional energy financial configuration mode, government-market dual-oriented regional energy financial configuration mode, and market-oriented regional energy financial configuration mode. The first configuration mode is suitable for areas with low energy finance levels. The second configuration mode is suitable for areas with a medium level of energy finance. The third configuration mode is suitable for areas with higher levels of energy finance [23]. Therefore, to study the energy financial allocation model applicable to each region, it is necessary to understand the energy financial level of each region. However, the level of energy finance is mainly reflected in the degree of consumption, so it is necessary to analyze the differentiated energy financial allocation models in different regions on the basis of obtaining energy consumption. Meanwhile, to obtain the energy consumption of each region, it is necessary to design an appropriate energy consumption prediction model to realize an effective prediction of the energy consumption of each region. Based on the prediction results, the energy financial level of each region is analyzed, and the differentiated configuration mode of energy finance applicable to each region is predicted. On this basis, the detailed features of such key influencing factors in each region are extracted by combining the attention mechanism. After its features are processed by feature fusion and normalization, they are used as input to the support vector machine prediction model, and the energy consumption of each area is output to realize the prediction of energy consumption in each area. Based on the prediction results, predict the differentiated allocation mode of energy finance in different regions.

A. Analysis of Key Factors Affecting Energy Consumption

The main factors that usually affect energy consumption include the level of urbanization, energy efficiency, demographic factors, industrial structure, and degree of economic development [24]. The demographic factors include two factors: population size and population structure. The degree of economic development includes factors such as the stage and scale of development. The structure of key influencing factors of energy consumption is shown in Fig. 1.

In Fig. 1, the effects of each key influencing factor on energy consumption are as follows.

1) Urbanization level factors: The difference in lifestyle between rural and urban residents has resulted in differentiated energy consumption in the two regions. Meanwhile, in the process of rural urbanization, the construction of transportation and infrastructure can also increase energy consumption. Therefore, the urbanization of the rural population has a promoting effect on energy consumption, and the proportion of the urban population can be used to measure the level of urbanization.

2) Energy efficiency factors: The improvement of energy efficiency can promote the sustainable development of energy, reduce energy waste, and effectively alleviate the contradiction between energy supply and demand, as well as achieve the purpose of saving energy and reducing emissions. The key indicator to measure energy efficiency is the value of energy consumption per unit of GDP, which is inversely proportional to energy efficiency.

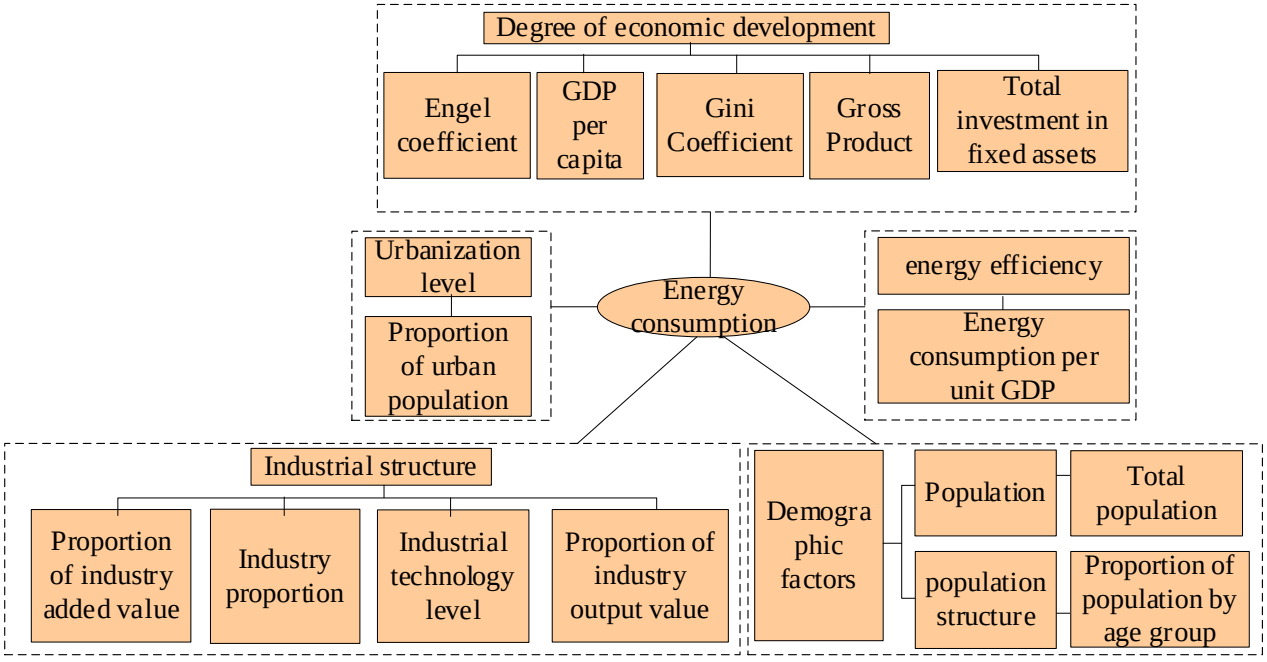


Fig. 1. Structure Diagram of Key Influencing Factors of Energy Consumption.

3) Demographic factors: With the increase in the number of people, the energy demand has gradually increased, so energy consumption has also shown an increasing trend. From the perspective of demographic structure, different age groups will lead to differences in consumption habits and lifestyles, as well as differences in energy consumption. Therefore, both population size and population structure will have an impact on energy consumption. Generally, the main indicators for measuring demographic factors are the total population, the proportion of the population of each age group, and so on.

4) Industrial structure factors: The differentiation of industrial structure has a direct impact on energy consumption, and the key energy consumption industry in my country is the secondary industry [25]. Common indicators used to measure changes in the secondary industry structure include the proportion of secondary industry's added value, the proportion of the secondary industry, the technical level of the secondary industry, and the proportion of secondary industry's output value. Therefore, the above indicators can also be used to measure the impact of changes in industrial structure on energy consumption.

5) Economic development degree factors: The increase in total energy consumption can promote the degree of economic development, and economic development will also bring about an increase in energy consumption. There is a mutually reinforcing relationship between energy consumption and economic development. Commonly used indicators to measure the degree of economic development include Engel's coefficient, GDP per capita, Gini coefficient, gross production value, and total fixed-asset investment. In which the Engel coefficient is inversely proportional to the living standards of the masses, and the Gini coefficient is directly proportional to the difference in income distribution, both of which can show the degree of economic development.

B. Feature Extraction of Key Influencing Factors for Energy Consumption based on Attention Mechanism

The attention mechanism is the process of extracting more detailed features. Its main principle is to quickly scan the overall scene of its visual interval through the human visual system, and select the key target interval from it through the brain's signal processing mechanism, and put more attention resources into this interval [26]. Feature extraction of factors affecting energy consumption is the key to energy consumption prediction. The attention mechanism can quickly extract more detailed feature information of the key influencing factors of energy consumption, which can effectively improve the accuracy and efficiency of the final energy consumption prediction in each region. The attention mechanism can search for many key-value pair mappings, and obtain the goodness of such key-value pairs. The goodness of fit is proportional to the number of allocated attention resources. The structure of the attention mechanism is shown in Fig. 2.

In Fig. 2, $L_i - W_i$ represents a key-value pair, which represents the weight of factors affecting energy consumption. The query is represented by Q, which represents energy consumption. The feature information of the key influencing factors of energy consumption extracted by the attention mechanism can be used to construct the prediction model, which can improve the prediction accuracy and efficiency of the model. The calculation process of attention is:

After calculating the goodness of fit between the energy consumption and the weight of each key influencing factor, the weights will be obtained. Here the splicing method is chosen to calculate the coincidence degree. The calculation equation is:

$$g\left(Q, L_i\right) = v_\alpha \left[Q; L_i\right] \tag{1}$$

In equation 1, the splicing coefficient is represented by v_α.

Step2: Normalization of weights. The weights obtained in step (1) are normalized by the $\mathrm{Soft\,max}$ function, and the calculation equation is:

$$\mathrm{Soft\,max}\left[g\left(Q, L_i\right)\right] = \left[\exp\left(g\left(Q, L_i\right)\right)\right] / \left[\sum_j \exp\left(g\left(Q, L_j\right)\right)\right] \tag{2}$$

Step3: Attention calculation. The weights gained by normalization of weighted and sum, and the weights acquired by factors of influencing energy consumption are used to obtain attention. The calculation equation is:

$$A\left(Q, L, W\right) = \sum_i \omega_i W_i \tag{3}$$

In equation 3, ω_i represents the normalized weight.

Based on the key influencing factors of energy consumption extracted by the attention mechanism, the detailed feature information is formed into a feature sample set. The sample set is processed accordingly and then input into the support vector machine prediction model to obtain the output of energy consumption in each region.

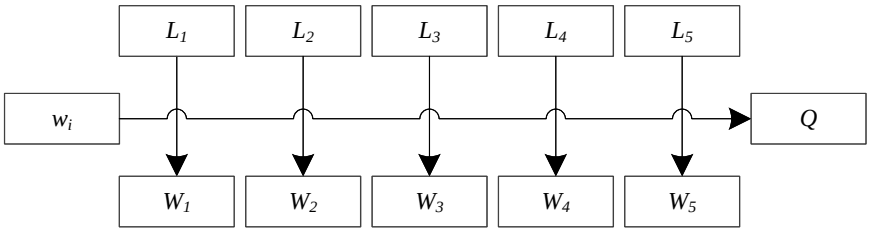


Fig. 2. Structure Diagram of Attention Mechanism.

C. Prediction of Regional Energy Consumption based on Extracted Feature Sample Sets

1) Feature fusion and normalization of the feature sample set: The feature sample set acquired by the attention mechanism can be fused to realize the complementary advantages of each factor feature in the sample set, improve the description performance of the feature, and help to further improve the prediction accuracy of the model [27]. After fusion, due to the dimensional differences of the characteristics of each factor in the feature sample set, it is easy to cause large fluctuations in the prediction results obtained by the prediction model, and the prediction performance is not stable enough. Therefore, it is necessary to further implement normalization processing on the basis of feature fusion to achieve stable prediction of the prediction model.

2) Feature fusion: Given that the features in the feature sample set obtained by the attention mechanism are the initial features, denoted by X_i *(i=1,2,...,M)*, where M represents the number of features. The features in the feature sample set after feature fusion processing are represented by X_j *(j=1,2,...,M)*. The feature fusion equation is:

$$X_j = \delta_j\left[\theta_i \times X_i\right] \tag{4}$$

In equation 4, δ_j represents the feature fusion function; the conversion function is represented by θ_i .

3) Normalization: After the fusion processing, the normalization of the feature data in the feature sample set is performed, that is, the value range of the feature data of each factor is unified to 0~1, so that can achieve the unity of dimensions and ensure the stability of the prediction results. The normalized processing equation is:

$$\tilde{x} = \left(x_i - x_{\min}\right) / \left(x_{\max} - x_{\min}\right) \tag{5}$$

In equation 5, $\tilde{x}$ represents the normalized feature sample set feature data; x_i represents the *i-th* feature data in the feature sample set after fusion processing. The highest value and the lowest value of the feature data in the feature sample set after fusion processing are represented by $x_{\max}$ and $x_{\min}$, respectively. Therefore, a new feature sample set $\tilde{X}$ composed of many normalized feature data $\tilde{x}$ can be obtained, which can be used as the input sample set of the prediction model to achieve a stable prediction of energy consumption in each region.

4) Support vector machine prediction model based on the processed feature sample set: Taking the processed new feature sample set XXX as input and the energy consumption of each area as output, a support vector machine prediction model is constructed to realize effective prediction of energy consumption in each area. The differentiated allocation mode of energy finance in each region is estimated by the prediction results.

Each feature $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_M$ in the input feature sample set $\tilde{X}$ is mapped to a high-dimensional feature space ($\sigma(X_1)$, $\sigma(X_2)$,…, $\sigma(X_n)$). According to statistics, the initial nonlinear model can be changed to the linear regression model of the high-dimensional feature space. Its equation is:

$$f\left(X_j\right) = d + \lambda^T \times \sigma\left(\tilde{X}_j\right) \tag{6}$$

In equation 6, *d* and λ represent the parameters that need to be identified in the linear regression model, where the adjustable weight vector is λ, and the bias is *d*. Based on the structural risk minimization criteria, the required identification parameters are processed, and the processing formula is:

$$H\left(f\right) = \sum_{i=1}^{s} \gamma \left\|\lambda\right\|^2 + B\left(e_i\right) \tag{7}$$

In equation 7, $B\left(e_i\right)$ represents the loss function; confidence risk is represented by $\left\|\lambda\right\|^2$; $H\left(f\right)$ represents empirical risk. According to the principle of support vector machine, the solution of formula (7) is equivalent to the operation of the following optimization problem, which is:

$$\begin{cases} \min K = B\sum_{i=1}^{s}\left(\mu_i + \mu_i^*\right) + \dfrac{1}{2}\lambda^T\lambda \\ s.t.\begin{cases} y - \left(\lambda, \sigma\left(X_t'\right)\right) - d \le \mu + \mu_i^* \\ \left(\lambda, \sigma\left(X_t'\right)\right) + d - y \le \mu + \mu_i^* \\ \mu_i^*, \mu_i \ge 0 \end{cases} \end{cases} \tag{8}$$

In equation 8, the optimized parameters are represented by μ_i and μ_i^* ; *K* represents the classification hyperplane; and the parameters of the inner product function are represented by X'_t.

By converting formula (8) into a dual problem and simplifying the calculation process, the nonlinear function obtained after conversion is expressed as:

$$f\left(X\right) = \sum_{i=1}^{k}\left(c_i - c_i^*\right)G\left(X_t', X'\right) + d \tag{9}$$

In equation 9, $G\left(X_t', X'\right)$ represents the inner product function; the support vector parameters are represented by c_i and c_i^*. The radial basis function is set by the Mercer condition, which is set the inner product function, and it is expressed as:

$$G\left(X_t', X'\right) = \exp\left\{-\left(\left\|\tilde{X}_j - \tilde{X}_a\right\|^2\right) / \varphi^2\right\} \tag{10}$$

In equation 10, φ^2 represents the Mercer condition coefficient; the training feature data vector and the test feature data vector in the feature sample set are represented by $\tilde{X}_j$ and $\tilde{X}_a$, respectively.

The equation (10) is substituted into equation (9). After equivalent transformation, it is acquired as:

$$f(X)=\sum_{i=1}^{l} c_j \exp\left\{-\left(\left\|\tilde{X}_j-\tilde{X}_a\right\|^2\right)/\varphi^2\right\}+d \tag{11}$$

In equation 11, c_j represents the parameter value corresponding to the support vector; the output vector set of energy consumption in each region is represented by $f(X)$. The energy consumption prediction parameters d *and* c_j can be acquired to obtain the energy consumption output of each region. The differentiated allocation mode of energy finance applicable to each region is predicted by the energy consumption of each region.

IV. Analysis of Experimental Results

Take a city as an example, the city is divided into three regions (a~c), and the method in this paper is used to study the differentiated configuration mode of energy finance in each region. Taking the period from 2011 to 2020 as an example, the historical statistics of energy consumption and key influencing factors in each area of the experimental city during the period are used as the basis. The proportion of the urban population, energy consumption per unit GDP, total population, the proportion of secondary industry added value, gross production value and energy consumption in each region of the city can be obtained, which is as Table I.

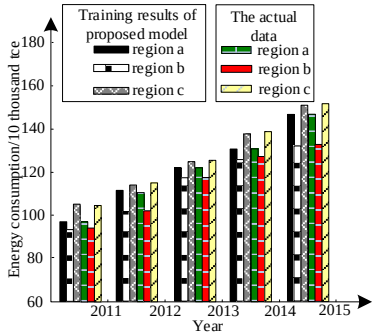
TABLE I. Historical Statistics of Energy Consumption and Various Influencing Factors in Each Region of the Experimental City from 2011 to 2020

Number	Year	Proportion of urban population /%	Energy consumption per unit GDP /tce/ 10 thousand yuan	Total population/10 thousand people	Proportion of added value of secondary industry /%	GDP/100 million yuan	Energy consumption/10 thousand tce
a	2011	48.8	4.7×10^{-5}	60.1	45.1	208.2	98.7
	2012	50.1	5.2×10^{-5}	60.8	44.8	211.5	110.3
	2013	52.4	5.6×10^{-5}	61.7	46.1	218.7	122.6
	2014	53.6	5.8×10^{-5}	62.5	45.8	225.4	130.7
	2015	54.8	6.3×10^{-5}	64.1	43.7	229.2	145.8
	2016	56	6.9×10^{-5}	65.7	41	233.4	162.4
	2017	57.9	7.3×10^{-5}	67.3	42.3	237.6	175.4
	2018	59.1	7.9×10^{-5}	69.2	44.5	241.7	191.1
	2019	59.8	8.2×10^{-5}	71	45.1	252.1	208.3
	2020	60.4	8.5×10^{-5}	72.5	42.5	260.4	221.6
b	2011	45.1	4.6×10^{-5}	55.7	43.7	199.5	93.2
	2012	46.4	5.1×10^{-5}	56.1	42.1	203.7	102.6
	2013	47.1	5.6×10^{-5}	56.9	45.1	208.9	118.5
	2014	47.8	5.8×10^{-5}	58.1	45.3	215.3	126.3
	2015	48.6	6.0×10^{-5}	59.2	41.4	220.7	133.8
	2016	50.1	6.1×10^{-5}	60.3	41.8	229.4	141.9
	2017	52	6.7×10^{-5}	62.7	43.8	233.6	156.9
	2018	53.8	7.1×10^{-5}	64.1	46.1	238.8	169.3
	2019	55.1	7.3×10^{-5}	65.8	44.3	246.4	180.4
	2020	56.7	8.1×10^{-5}	68	43.2	252.2	206.2
c	2011	50.6	4.9×10^{-5}	65.2	46.3	212.4	105.2
	2012	52	5.3×10^{-5}	66.7	45.5	220.6	117.3
	2013	52.8	5.5×10^{-5}	68.1	46.2	227.8	126.9
	2014	54.1	5.8×10^{-5}	69.7	44.1	236.3	138.5
	2015	55	6.2×10^{-5}	71.2	43.7	244.2	151.3
	2016	56.3	6.4×10^{-5}	72.8	47.1	251.7	162.4
	2017	57.8	6.8×10^{-5}	74.1	46.8	258.9	176.9
	2018	59.1	7.2×10^{-5}	76.3	42.1	267.3	192.6
	2019	60.5	7.7×10^{-5}	77.9	46.1	274.2	211.1
	2020	61.3	7.9×10^{-5}	79.5	45.8	285.2	226.9

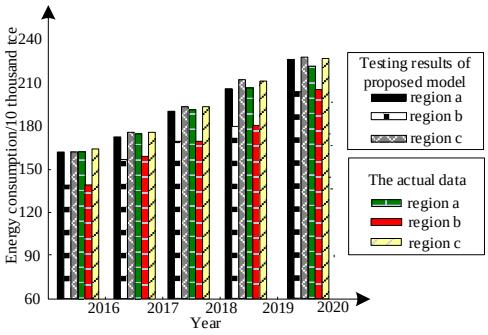
Using historical statistical data from 2011 to 2015 as the training sample set, and historical data from 2016 to 2020 as the test sample set, the prediction model of this paper is used to implement training and testing, which can obtain the training effect and testing effect, as shown in Fig. 3.

It can be seen from Fig. 3 that the training results and testing results of the prediction model of the method in this paper are in close agreement with the actual energy consumption in historical statistics, and the maximum error value does not exceed 15,000 tce. Therefore, the training effect and test effect of the prediction model in this paper are relatively ideal, and the obtained prediction results have high accuracy, which can be applied to the actual energy consumption prediction in the study of the differential configuration mode of energy finance for different regions.

This paper studies the differentiated energy financial allocation modes of the experimental cities from 2022 to 2030 in the future, and further uses the proposed prediction model to predict the energy consumption of the experimental cities from 2022 to 2030. The future energy financial differentiated allocation mode of each region is studied by the obtained prediction results. The prediction results of energy consumption in each region of the experimental city from 2022 to 2030 are shown in Fig. 4.



(a) The Training Effect of the Prediction Model for the Proposed Method by the Paper.



(b) The Testing effect of the Prediction Model for the Proposed Method by the Paper.

Fig. 3. The Training and Testing Effect of the Prediction Model in the Method of this Paper.

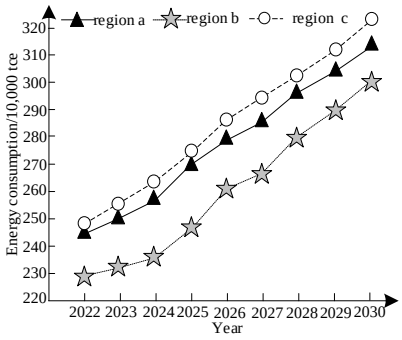


Fig. 4. The Prediction Results of Energy Consumption in each Region of the Experimental City from 2022 to 2030.

From Fig. 4, it can be concluded that the annual energy consumption of each region for the experimental city from 2022 to 2030 will show an upward trend. On the whole, energy consumption in region *c* is higher, energy consumption in the region *a* is at a medium level, and energy consumption in region *b* is the lowest. Combining this prediction result, the future energy financial level of the three regions of the experimental city is ranked *c*-*a*-*b* from high to low. Therefore, the energy financial allocation models corresponding to each region from 2022 to 2030 in the future should be: region-c-market-oriented, region-a-government-market of dual-oriented, and region-b-government-oriented. Based on the predicted different regional energy finance allocation models in the future, it can provide a clear development direction for the future development of energy finance in each area of the experimental city. It also improves the future development level of energy finance and provides help to effectively avoid bottlenecks in future development.

V. Conclusion

This paper focuses on the research on the prediction method of regional differential allocation pattern of energy finance based on attention mechanism and support vector machine. It extracts many detailed features of key influencing factors of energy consumption by using attention mechanism, and after fusing and normalizing them, the feature sample set obtained is input into the support vector machine, constructs the support vector machine prediction model, and outputs the prediction results of regional energy consumption. According to this result, the energy finance allocation mode applicable to each region is analyzed. The proposed prediction model has high training prediction accuracy and test prediction accuracy, and has a good prediction effect on the historical sample set. It can provide a scientific guidance path for the sustainable and stable development of energy finance in various regions of the city in the future.

Acknowledgment

The study was supported by “Philosophy and social science research project of NIIT (Grant No. 2019SKYJ02)”.

REFERENCES

[1] F. E. A. Mills, J. Dong, L. Yiling, M. A. Baafi, B. Li and K. Zeng, "Towards sustainable competitiveness: How does financial development affect dynamic energy efficiency in Belt & Road economies?," Sustainable Production and Consumption, 2021, 27, pp. 587-601.

[2] C. Zhang, J. Fu and Z. Pu, "A study of the petroleum trade network of countries along The Belt and Road Initiative," Journal of Cleaner Production, 2019, 222, pp. 593-605.

[3] W. E. I. Li, "The Threshold Effect of Clean Energy Development on Employment Based on the Perspective of Financial Scale," Economic Theory and Business Management, 2021, 41(9), pp. 99.

[4] S. Saud and S. Chen, "An empirical analysis of financial development and energy demand: establishing the role of globalization," Environmental Science and Pollution Research, 2018, 25(24), pp. 24326-24337.

[5] H. Fan, Q. Q. Yuan and J. Deng, "Two-stage capacity optimization configuration method for multi-region integrated energy system," Modern Electric Power, 2020, 37(5), pp. 441-449.

[6] Z. LIU and P. WAN, "Pedestrian re-identification feature extraction method based on attention mechanism," Journal of Computer Applications, 2020, 40(3), pp. 672-676.

[7] X. Qin, J. Jiang, C. A. Yuan, S. Qiao and W. Fan, "Arbitrary shape natural scene text detection method based on soft attention mechanism and dilated convolution," IEEE Access, 2020, 8, 122685-122694.

[8] F. Kaytez, "A hybrid approach based on autoregressive integrated moving average and least-square support vector machine for long-term forecasting of net electricity consumption," Energy, 2020, 197, pp. 117200.

[9] D. Koschwitz, J. Frisch and C. Van Treeck, "Data-driven heating and cooling load predictions for non-residential buildings based on support vector machine regression and NARX Recurrent Neural Network: A comparative study on district scale," Energy, 2018, 165, pp. 134-142.

[10] M. S. Islam, A. Q. Al-Amin and M. S. K. Sarkar, "Energy crisis in Bangladesh: Challenges, progress, and prospects for alternative energy resources," Utilities Policy, 2021, 71, pp. 101221.

[11] W. Bishan, "Development mechanism of energy industry cluster in new normal economy," Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2019, 41(23), pp. 2853-2860.

[12] X. Zhong and J. Tu, "Study on Energy Finance Risk Warning Model---Based on GABP Algorithm," In 2017 3rd International Conference on Economics, Social Science, Arts, Education and Management Engineering (ESSAEME 2017). Atlantis Press (2017, July).

[13] D. Y. Peng, "Empirical Study on Coupling and Coordinated Development of"Energy-Finance-Environment" in the Middle Reaches of Yangtze River," Proceedings of 2019 4th International Conference on Advances in Energy and Environment Research (ICAEER 2019), 2019, pp. 912-917.

[14] J. Puck and I. Filatotchev, "Finance and the multinational company: Building bridges between finance and global strategy research," Global Strategy Journal, 2020, 10(4), pp. 655-675.

[15] O. L. A. N. I. Y. I. Evans and O. R. Alenoghena, "Financial inclusion and GDP per capital in Africa: A Bayesian VAR model," Journal of Economics & Sustainable Development, 2017, 8(18), pp. 44-57.

[16] Y. Ma, R. Han and X. Fu, "Stock prediction based on random forest and LSTM neural network. In 2019 19th International Conference on Control," Automation and Systems (ICCAS) IEEE. 126-130 (2019, October).

[17] S. Ma, H. Wang, B. Xu , H. Xiao, F. Xie, H. N. Dai and T. Wang, "Banking Comprehensive Risk Management System Based on Big Data Architecture of Hybrid Processing Engines and Databases," In 2018 IEEE Smart World, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smart World/SCALCOM/UIC/ATC/CBD Com/IOP/SCI) IEEE. 1844-1851 (2018, October).

[18] S. Li, L. Yuan and L. Wenjun, "China's Regional Financial Risk Spatial Correlation Network and Regional Contagion Effect: 2009-2016," Management Review, 2019, 31(8), pp. 35.

[19] R. Shaidullin, E. Bulatova, L. Kurmanova, R. Khabibullin and J. Zhuzhoma, "Evaluation of financial stability of Russian companies," In E3S Web of Conferences (Vol. 110, p. 02044). EDP Sciences (2019).

[20] W. Dariusz, "Financial geography II: The impacts of FinTech – Financial sector and centres, regulation and stability, inclusion and governance," Progress in Human Geography, 2021, 45(4).

[21] M. G. Caldas, V. Matheus, D. M. C. Oliveira, "Impacts of the sovereign risk perception on financial stability: Evidence from Brazil," Quarterly Review of Economics and Finance, 2021, 81.

[22] D. Gao, G. Li, Y. Li and K. Gao, "Does FDI improve green total factor energy efficiency under heterogeneous environmental regulation? Evidence from China," Environmental Science and Pollution Research, 2021, 1-14.

[23] G. Shi, D. Liu and Q. Wei, "Energy consumption prediction of office buildings based on echo state networks," Neurocomputing, 2016, 216, pp. 478-488.

[24] X. Y. Zhou and A. L. Gu, "Impacts of household living consumption on energy use and carbon emissions in China based on the input–output model," Advances in Climate Change Research, 2020, 11(2), pp. 118-130.

[25] D. Fang and B. Yu, "Driving mechanism and decoupling effect of PM2. 5 emissions: Empirical evidence from China's industrial sector," Energy Policy, 2021, 149, pp. 112017.

[26] H. Wang, J. Shi and Z. Zhang, "Semantic relation extraction of LSTM based on attention mechanism," Computer application research, 2018, 35(5), pp. 143-146.

[27] H. S. Shin, H. Y. Kwon and S. J. Ryu, "A new text classification model based on contrastive word embedding for detecting cybersecurity intelligence in twitter," Electronics, 2020, 9(9), pp. 1527.

Bilingual AI-Driven Chatbot for Academic Advising

Ghazala Bilquise[1]
Computer Information Science Department, Higher Colleges of Technology, Dubai
United Arab Emirates

Samar Ibrahim[2]
School of Arts and Sciences
American University in Dubai
Dubai, United Arab Emirates

Khaled Shaalan[3]
The British University in Dubai
Informatics Department
The British University in Dubai

***Abstract*—Conversational technologies are revolutionizing how organizations communicate with people, thereby raising quick responses and constant availability expectations. Students often have queries about the institutional and academic policies and procedures, academic progression, activities, and more in an academic environment. In reality, the student services team and the academic advisors are overwhelmed with several queries that they cannot provide instant responses to, resulting in dissatisfaction with services. Our study leverages Artificial Intelligence and Natural Language processing technologies to build a bilingual chatbot that interacts with students in the English and Arabic languages. The conversational agent is built in Python and designed for students to support advising-related queries. We use a purpose-built domain-specific corpus consisting of the common questions advisors receive from students and their responses as the chatbots knowledge base. The chatbot engine determines the user intent by processing the input and retrieves the most appropriate response that matches the intent with an accuracy of 80% in English and 75% in Arabic. We also evaluated the chatbot interface by conducting field experiments with students to test the accuracy of the chatbot with real-time input and test the application interface.**

Keywords*—Chatbot; conversational agent; academic advising; natural language processing; deep learning; bilingual English Arabic

I. Introduction

Conversational technologies are transforming the interaction landscape between organizations and people, causing digital communication to be propelled by technology rather than humans. A chatbot, also known as a conversational agent, is a software system that processes and simulates human conversation to provide digital assistance in real-time [1]. The constant availability of chatbots and the ability to respond immediately and communicate in a natural language have escalated their popularity across all domains [2], [3]. Chatbots are being entrusted with various tasks previously handled by human agents, such as providing customer service, healthcare advice, e-shopping, and answering queries. Organizations pervasively rely on chatbots to support customers' needs and increase customer satisfaction with services. Therefore in this digital era, chatbots have the potential to support student queries and assist in the academic advising process in the education domain.

Academic advising is an integral function of Higher Education Institutions (HEIs) and has been widely acknowledged as a principal strategy for confronting the challenges of persistence and retention [4]–[6]. While advising encompasses several tasks, one of the crucial tasks of advising is to provide students with the essential information required for navigating their academic journey successfully. This task involves a high degree of interaction between advisors and students and often leads to dissatisfaction with advising services when students cannot get timely and accurate information. The large number of students assigned to each advisor makes it impossible for the advisor to respond to all students in a satisfactory amount of time [7]. Moreover, students' expectations and information requirements for their daily tasks have intensified with today's technological advancement. Providing adequate channels for student communication is vital for their academic progression and integration with the academic environment. Therefore, a chatbot can provide numerous benefits to the students and the academic institution by providing instant responses to students, thereby enhancing student satisfaction.

This study aims at building a chatbot for the students at an academic institution in the UAE. The institution offers four undergraduate programs of study. There are nearly 3000 students of Arab origin and almost 100 faculty members employed at the institution. Each faculty member serves as an advisor to nearly 25-30 students per semester. This large ratio makes it challenging for the advisor to contribute quality time to advisees and answer all their queries or make them aware of the college policies related to registration, courses, pre-requisites, and more. A chatbot would assist in reducing the workload of the advisor so they may focus on more cognitive tasks such as creating an ideal study plan for their advisees.

Considering the aforementioned challenges of advising at the institution of study, the study aims to develop a chatbot that supports students in answering queries on college and academic-related matters and thereby improve student satisfaction with college services. The chatbot will be bilingual and provide an interface in both English and Arabic. Moreover, the chatbot will be developed using a neural network and Natural Language (NLP) technologies. Thus, our study is novel in its context with bilingual conversational support.

The rest of the paper is organized as follows. Section II provides a literature review on the background of chatbots and related studies of chatbot use in the education sector and bi-lingual chatbots. Next, Section III describes our research methodology, while Section IV presents the evaluation and results of the study. Finally, the study concludes with Section V, which summarizes the paper, significance of the study, limitations, and new directions for future research.

II. Literature Review

A. Overview of Chatbots

Chatbots, are dialog systems that mimic human conversations in text, voice, or multimodal form [1]. A chatbot, also known as a conversational agent, processes user input to discover the query's intent and respond appropriately. In the last few years, there has been a tremendous rise in chatbot applications worldwide [3]. Organizations rely on chatbots to respond to customer service queries and automate tasks [8]. Chatbots are also being used in the healthcare sector for psychiatric and medical diagnosis, raising awareness of health and safety issues [9], [10]. In the educational sector, chatbots are used for teaching and learning activities, student advising, and administrative tasks [11]. Chatbots offer a cost-effective means of delivering services to consumers eliminating repetitive and time-consuming human-agent communication, enabling them to focus on high-end complex tasks [2].

Several classifications exist in literature to categorize chatbots. A chatbot may be rule-based or driven by Artificial Intelligence (AI). A rule-based chatbot provides predefined responses based on keywords and a defined set of rules. ELIZA and PARRY were among the earliest rule-based chatbots developed in the 1960s, built using pattern recognition technology [3]. Artificial Intelligence Markup Language (AIML) [12] was used to develop the ALICE chatbot in 1995. The markup language is based on an XML structure. Chatbots developed with AIML use a rule-based approach to respond to user queries based on inputs that match a pattern.

On the other hand, an AI-driven chatbot is powered by NLP techniques to recognize the intents of the user input and generate an appropriate response based on the intent. AI-driven chatbots are technologically superior and can meet consumers' language and conversational expectations [3]. Several AI techniques have been employed in the literature to develop chatbots, such as machine learning, neural network [13], deep learning with sequence to sequence model [14], and CVAE Models [15].

Chatbots have been classified as task-oriented or non-task-oriented based on their functionality [16]. A task-oriented chatbot responds to domain-specific user queries and performs tasks such as making a reservation, placing an order, or answering queries. On the other hand, a non-task-oriented chatbot responds to open-ended queries that are not domain-specific, also called an open-domain. The main purpose of these chatbots is to act as digital assistants using an open-ended dialog. Siri and Alexa are an example of non-task-oriented virtual assistants.

Chatbots have also been classified based on their response generation method as retrieval-based and generative chatbots [17]. A retrieval-based chatbot retrieves responses from a knowledge base using machine learning algorithms, and NLP techniques process the user input, allowing users to communicate in natural language. However, the responses generated in a retrieval-based chatbot are fixed. On the other hand, a generative chatbot is trained on a conversational corpus to generate new and diverse responses that do not exist in the dataset. A limitation of the generative model is that it requires massive training data and may provide unpredictable responses not stored in the corpus.

This study uses a domain-specific knowledge base to develop a task-oriented chatbot that responds to student queries. The students ask questions in a natural language, yet the responses provided by the chatbot must be precise and accurate. Hence we use an AI-driven retrieval-based chatbot that uses NLP techniques to process user input and retrieve precise responses from a corpus of advising queries. The chatbot determines the user intent by processing the input and retrieving the response that matches the intent.

B. Chatbots in Education

Some studies used NLP techniques with a rule-based approach for developing chatbots in the educational setting to answer student queries [18], [19]. Reference [18] developed a rule-based conversational agent using PHP and NLP to respond to student queries with an accuracy of 80%. While reference [19] developed a chatbot using a social conversation dataset between students and advisors. The chatbot was developed using a frequent intent pattern by discovering rules from the dataset.

Several studies develop retrieval-based chatbots to answer student queries using AI and NLP techniques. Refrence [20] developed a chatbot based on pattern matching using AIML and Latent Semantic Analysis (LSA). The chatbot answers informational queries on college and academics. In a similar study, [21] proposed a chatbot that answers frequently asked questions. The knowledge base of the chatbot consisted of 300 questions. Both studies did not evaluate the performance of the chatbot.

Reference [22] developed an AI-driven chatbot that allows students to enquire about college admission rules and policies. The chatbot is developed using the RASA framework. The performance of the chatbot was evaluated using the confidence of the responses. However, the confidence does not indicate the accuracy of the response. Moreover, the study did not specify how they handled spelling errors in the user input.

In another study, [13] developed a chatbot using machine learning and NLP techniques that answer campus-related queries published as FAQs on the website. The study compares two chatbot models, RNN based Seq2Seq model and a semantic similarity model. The results show that the semantic similarity model performs better in cases where the dataset size is small. Furthermore, this study uses a deep learning model to process the input patterns and retrieve the most accurate response rather than constructing responses, similar to our study. However, the chatbot is developed in one language only.

Several studies have developed chatbots to answer students' admissions, policies, or academic advising queries. However, only a few have used neural networks with NLP techniques to process the user input.

C. Chatbots in Arabic and other Languages

Due to its complexity, the Arabic language is underrepresented in NLP and chatbot development and is not given enough attention by researchers. Few studies have

examined Arabic chatbots in general and education, some of which were bilingual or multilingual.

BOTTA is an Arabic Egyptian dialect female public chatbot proposed by [23] that simulates friendly conversations with users. It is a retrieval-based model designed for open domain conversations responds. Arabchat and enhanced ArabChat are conversational agents designed for students at Applied Science University in Jordan [24]. Both are interactive chatbots that use Arabic MSA textual language. The study [25] proposed a conversational social chatbot "Nabiha" for Information Technology (IT) students at King Saud University using the Saudi Arabic dialect. Nabiha is a retrieval-based chatbot that uses AIML. It serves as an academic counselor to interact with students about their courses and academic progress inquiries.

A bilingual chatbot called “Jooka” was designed by [26] to improve the admissions process at the German University in Cairo (GUC). It understands queries written in English and Arabic and responds based on the query language. Google Cloud, Translation API was used to translate Arabic to English. However, in our study we found that translation of APIs for Arabic language are still not mature and result in an unnatural response.

Reference [27] proposed a voice-interactive chatbot that adopts a multilingual interface specifically to detect and respond to exam stress of university students. The chatbot application analyzes the tone of the user's voice to determine their feelings towards their exams with an accracy of 76.5%.

Multilingual chatbots have been developed in domains other than education. For example, [28] proposed a multilingual health chatbot application that can diagnose disease based on user symptoms and supports three languages: English, Hindi, and Gujarati. Reference [29] presents a bilingual retail chatbot that can handle Filipino-Tagalog and English languages that employs k-fold cross-validation on a dataset generated using a bilingual automatic corpus engine.

Supporting users with chatbot conversations in English as well as local dialects is gaining importance and is highlighted in the literature. The above studies show that there are two ways of creating a bilingual chatbot. The first method is using translation services to perform translation between both languages while maintaining a corpus in the primary language only. And the other method is to create and maintain a two corpus files, one for each language. Our study adopts the second method as experimentation with the first method resulted in unnatural translation between English and Arabic languages.

III. Methodology

This section presents the methodology adopted for planning, designing, and developing the chatbot system. This system provides bilingual advice through a chatbot with an easy-to-use interface to communicate in either Arabic or English. Our chatbot is equipped with sufficient information to provide students with answers to their specific advising inquiries. The advice chatbot adopts a bilingual corpus as the knowledge source type used to generate responses adopting the retrieval-based model. As part of the retrieval-based model, a chatbot uses heuristics to select the most appropriate response from a predefined pool of responses. This retrieval-based model is selected due to the need for precise and accurate responses to a specific task and domain. The following sub-sections present the three phases of the methodology - data collection, building the chatbot model, and the chatbot GUI development.

A. Data Collection

The conversational data required for the chatbot was collected through interaction with students, advisors, and referrals to university policy documents. The data consists of the most commonly asked queries that advisors usually receive from students and responses to those questions. We followed four main steps in collecting the conversational data required for the chatbot. First, we identified eight primary contexts to classify each query. The context is the domain of the user’s request, such as attendance, course delivery, and more. Second, we added queries to the contents and tagged each query with a unique intent tag that identifies the main purpose of the query. Third, we created patterns for each query to depict the variety of ways the question may be presented to the chatbot. Last, we added a variety of responses for each intent to incorporate diversity in the response.

In summary, each intent reflects what students would like to accomplish when interacting with the chatbot. Table I illustrates the different contexts and the number of intents in each context. For the purpose of this study, we developed 152 English and Arabic intents, with a total of 356 patterns.

TABLE I. Distribution of Intents and Patterns

Context	Number of intents	Patterns	Description
Greeting	8	28	That greet, welcome, and thank the user
Academic Standing	22	50	Students’ academic status/probation
Registration	32	76	inquiries related to registration/ scheduling and retaking courses
Summer	6	24	Inquiries related to Summer Courses /credits
WP	26	56	Inquiries related to work placement, schedule/registering.
COVID	14	28	Inquiries related to requirements related to COVID on campus
Final Exam	20	48	Inquiries related to materials scheduling /attending / missing / to final exams
Attendance	6	12	Inquiries related to attendance
Course Delivery	18	34	Inquiries about online, Hybrid courses
	152	356	

The corpus, consisting of the conversational data, was stored in JSON format. We use two corpus files to store the English intents and the other to store the Arabic intents as the initial experiments revealed that translation services from Arabic to English and vice versa are still very weak and result in unnatural statements. For example, when translating the Arabic statement “ما هو المعدل المناسب للنجاح”, the resulting translation is “what is the appropriate rate of success,” which is

not a natural way of phrasing the statement in the English language.

The complexity and NLP challenges inherent in the Arabic language, such as dialectal differences, orthographic ambiguity, and inconsistencies, are more prevalent while translating [23]. Moreover, the existing translation functions are inaccurate and do not reflect the correct English statements.

Furthermore, using a separate Arabic corpus allows us to integrate English words that are typically used by students when they write in the Arabic language, such as “probation”, “covid,” “GPA,” and more. Also, the Arabic corpus uses Arabic words written in local dialects. Table II shows sample intents from the Arabic and English corpora, with patterns, and responses.

After building the chatbot, we conducted a pilot implementation with eight students and three faculty members to augment the corpus with additional queries. Students and advisors were asked to type questions in natural language (English and Arabic) within the contexts identified earlier. The purpose of the pilot was to re-examine the initial corpus and augment it with additional patterns in which a query may be composed by the user. Furthermore, the pilot was also meant to identify any gaps in data collection within the contexts identified. After conducting the pilot, we examined the results and added new intents or patterns to existing intents. In addition, we identified queries that were not addressed in the initial corpus development; for example, questions, such as blackboard password, arriving before the final exam, and materials needed for the final exam were not included in the initial corpus development. Therefore, the pilot implementation was crucial to extend the corpus.

B. Chatbot Modelling with Deep Learning

The chatbot model was developed in Python using a supervised deep learning algorithm. Deep learning is a subset of machine learning based on an artificial neural network, in which layers of nodes simulate the neurons of a human brain. Input neurons are interconnected with multiple hidden layers to produce output by automatically adjusting the weights of the nodes in each layer [30]. We used the keras library in Python to build our deep learning network to build two chatbot models, each trained on the English and Arabic corpus, respectively. Fig. 1 shows the steps involved in developing the English chatbot model. Similar steps were also applied for developing the Arabic chatbot model.

First, to train our chatbot model, we pre-processed the training data and encoded each intent to make it suitable for the neural network algorithm. Pre-processing is crucial to transform the corpus data in an appropriate form for the neural network algorithm. Pre-processing the data enhances the efficiency and performance of the model. The pre-processing phase includes transforming input to lower case, removing punctuations and special characters, tokenization, and vectorization of the words. We used the NLTK library in Python to perform all the pre-processing steps.

Tokenization is the process of extracting words from sentences. We tokenized all the patterns in the corpus to extract individual words. The words were then simplified to their base forms using the process of lemmatization and stemming for English and Arabic words, respectively. Lemmatization converts the words to mean their original form based on the context, while stemming reduces the words to their base by removing the last characters without preserving the meaning. We used the NLTK WordNetLemmatizer library to lemmatize the English words using the parts of speech tag. For the lack of a sound library in Python for lemmatizing Arabic words, we used the ISRIStemmer to stem the words. However, some Arabic words did not retain their meaning when stemmed, such as "يمك" does not have any meaning originally "يمكنني," also “ادة" did not preserve its meaning originally “مادة.” Table III shows input patterns in Arabic and English and the extracted words. Word extraction and reducing to its base form resulted in 247 unique words in English and 250 words in Arabic.

TABLE II. SAMPLE ARABIC AND ENGLISH INTENTS

Description	Sample Intent
Arabic intent where there is the use of a UAE dialect such as “أقدر”	{“tag”: “numberofcurses-ar,” "patterns": [" كم عدد المواد أقدر اسجل ", " كم مادة أقدر اسجل","كم مادة ممكن اسجل"], "responses": ["هذا يعتمد على مستواك الاكاديمي"], "context_set": “academic-standing”, }
Arabic intent that includes English word “Probation” and written in Arabic ” بروبيشن” and also using a dialect UAE “شتقصد” and using	{“tag”: “Probation-ar,” "patterns": [" Probation ما معنى ", "بروبيشن ما معنى تحت الاختبار","شتقصد"], "responses": ["هذا يعتمد على مستواك الاكاديمي"], "context_set": “academic-standing”, }
English intent for the registration context student clarifying of a section that is full in a course	{"tag": "sectionfull-en", "patterns": ["If all sections are full", "There are no seats available"], "responses": ["Contact your advisor to change your plan"], "context_set": "registration",}

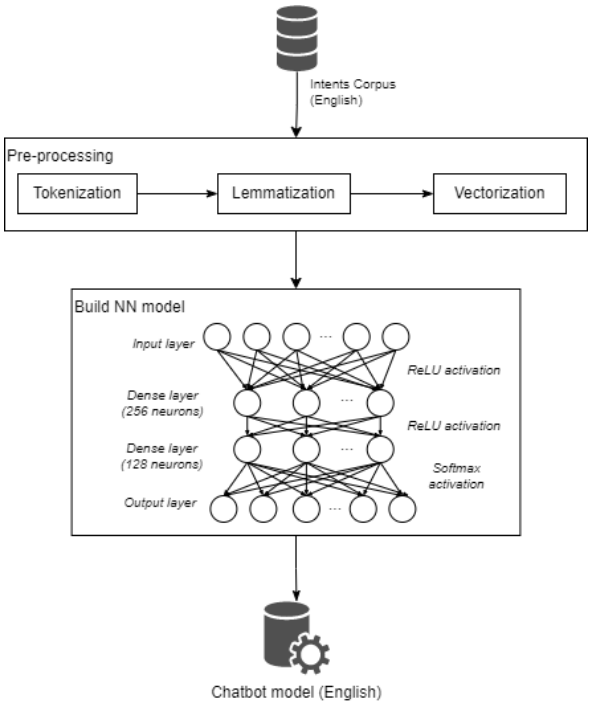


Fig. 1. English Chatbot Model Development Process.

TABLE III. SAMPLE OF LEMMATIZATION AND STEMMING OF PATTERNS

Pattern	Extracted words
‘What is going to happen if I do not raise my GPA	['what', 'be', 'go', 'to', 'happen', 'if', 'I', 'do', 'not', 'raise', 'my', 'gpa']
‘What is the time for adding and dropping courses’	['what', 'be', 'the', 'time', 'for', 'add', 'and', 'drop', 'course']
كيف يمكنني معرفة ما إذا كانت المادة عبر الإنترنت أم لا	['كيف', 'يمك', 'عرف', 'ما', 'اذا', 'كانت', 'ادة', ' "عبر", 'ترن', 'ام', 'لا]
ماذا أفعل بحضوري إذا لازم أذهب للخدمة العسكرية	['اذا', 'فعل', 'حضر', 'اذا', 'لزم', 'ذهب', 'خدم', ' "عسكر]

The next step of pre-processing is the process of vectorization. In this step, the words were converted to numerical form by creating a list of word vectors, which is a two-dimensional representation of each unique word and its frequency of occurrence. These word vectors are used as features of the neural network input layer.

After the pre-processing phase, we build two Neural Network (NN) models with deep learning for English and Arabic, respectively. The keras library in Python was used to build the NN model. The network consists of an input layer, two hidden layers, also known as the dense layers, and the output layer. The input layer comprises of all the unique features extracted from the respective corpus and has approximately 250 neurons in each model. The output layer represents the classes or the intents that should be predicted.

The first dense layer has 256 neurons, and the second has 128 neurons with a dropout rate of 0.5. The number of neurons in the layers is considered ideal since a smaller number would lead to underfitting, and a larger number would result in overfitting. Therefore, we selected the number of neurons in the dense layers between the input and output neurons. We configured the neural network with the following settings:

Optimizer – Stochastic Gradient Descent (SGD). The SGD estimates the expected risk gradient based on a single randomly selected sample instead of computing the precise value. Thus it is an optimization algorithm because the samples are randomly selected from the distribution [31].

Activation Function – Rectified Linear Unit (ReLU), was used as an activation function in the hidden layers. ReLU is a piecewise function in which if the input value is zero or less, then the output value will also be zero; otherwise, the output value will equal the input value. When data value is forced to be zero, a sparse characteristic is created, making the function fast and efficient. In addition to providing a faster computer rate, the ReLU function does not cause gradient diffusion problems, i.e., minor errors. However, because it always returns 0 for negative values, it can kill some neurons permanently and affect the final results or the output, i.e., generate exploding gradients [32].

Learning Rate – 0.01. The learning rate is a configurable hyperparameter used in training neural networks. Typically, between 0.0 and 1.0, it has a small positive value that must be carefully selected. That value determines how quickly the models are adapted to the problem. Lower learning rates result in more training cycles, and the process can get stuck, whereas larger learning rates lead to rapid changes and require fewer training cycles [33].

Classification function – Softmax. In artificial neural networks, the classification function, also known as the activation function, identifies a node's output given an input or set of inputs. The activation function allows neural networks to recognize complex relationships and patterns in data. This refers to the activated neurons features that can be retained and mapped out by nonlinear functions and employed to solve complex nonlinear problems. Furthermore, the activation function increases the neural network’s ability in which the nonlinear ability of the activation function makes the deep neural network have real artificial intelligence [32].

Epoch – 200. The epoch determines how many cycles are used to train the model. Since the dataset size is small, we set the epoch size to 200.

C. Chatbot Engine and GUI

The chatbot engine interacts with the Graphical User (GUI) to get the user query as an input and returns the most suitable response. Fig. 2 shows the architecture of our chatbot engine. There are three logical components of our chatbot engine – Natural Language Understanding (NLU), Natural Language Processing (NLP), and Natural Language Generation (NLG).

1) NLP: In this component, the user submits a query; the chatbot application first determines the language used for communication and accordingly uses the appropriate chatbot model for getting the response. The input query is first corrected for spelling mistakes using a spell check function implemented from the TextBlob Python library. In the case of Arabic input, however, there is no spell check function performed due to the complexity of the Arabic language and the inconsistency of the spellchecking function on the Arabic language, which led to many errors while applying it. This is considered a limitation of the study and a potential area for further development, research, and analysis. Also, in the NLP component, the input query is pre-processed using the same methods used in the training phase: tokenization, lemmatization/stemming, and vectorization of the words. In addition, all intents features are extracted from the input query in this component.

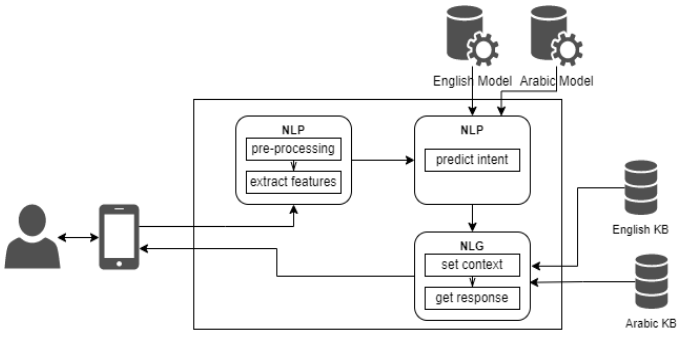


Fig. 2. Chatbot Engine.

2) NLU: This component bridges the gap between what computers understand and how people speak. The appropriate chatbot model is used for prediction by providing the word vectors to the two different models, Arabic or English, for classification. The prediction returns all the matching intents along with the probability of prediction. We set an error threshold of 25% to accept all predictions that have a probability above this threshold. Thus, in this component, if the model is not confident of the intent it detects, the user is requested to rephrase or restate their intent because of missing vocabulary.

3) NLG: In this component, the user's intent context is set based on the user query and language selected. The prediction is performed according to the training model discussed in the previous section. The function matches the intents tags and generates the response from either the Arabic or English knowledge base. If the model is unable to generate the response, a message will be displayed in English or Arabic, "contact your advisor," "اتصل بمشرفك."

The advising chatbots Graphical User Interface (GUI) was developed using Python's tkinter library. Our chatbot application employs a simple natural language user interface similar to an instant messaging application, which has a text box to type the input, a button to submit the message, and a display to show the input and response of the chat conversation. In addition, our interface consists of a language button that allows the user to toggle between the English or Arabic language mode to communicate with the chatbot. Fig. 3 shows three screenshots of an English and Arabic conversation, respectively. The screenshot (a) shows that the chatbot accepts spelling errors as the spellings are corrected in the pre-processing phase. For example, despite the spelling mistake of the word "available," the chatbot retrieves the correct response. In screenshots (b), the chatbot appends an additional message to rephrase the question when the response retrieval has a low confidence rate (below 0.75).

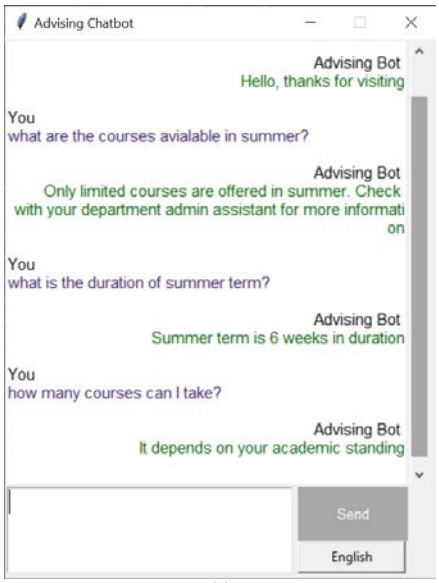
IV. Evaluation and Results

Evaluation metrics are essential to determine the machine learning algorithm's performance and assess the chatbot application. Since there are no standard evaluation methods of a chatbot application [34], the evaluation measure should be adapted to the chatbot type of service. Some studies used both automatic and human evaluations to measure the performance of chatbots [35].

Automatic evaluation measures the machine learning model's performance using known metrics such as accuracy, F1-Score, BLEU, and more, while human evaluation measures the quality of the responses using people as evaluators. Hence, human evaluation is suitable for generative chatbots that generate diverse responses, which do not exist in the corpus. However, since our chatbot is retrieval-based, we use only automatic evaluation to assess the chatbot performance.

We used two methods of evaluation. First, we used a test set consisting of queries with labeled intents, and second, we used human input to test the chatbot application and performance using ad-hoc queries. Finally, we used accuracy as a metric to evaluate the chatbot model for both methods. Accuracy measures the ratio of correct responses over the total responses that are predicted on an unlabeled set of inputs.

We developed two new test sets in the English and Arabic languages in the first evaluation method, which were not used to train the model. Each set is approximately 30% of the corpus size. The test set is populated with queries labeled with the actual intent tags. The label is hidden from the prediction algorithm when the test is performed. Table IV shows a few sample queries in English and Arabic from the test set. The queries are phrased differently than the patterns that exist in the corpus. The evaluation aims to determine the percentage of accurate responses retrieved by the chatbot. After running the prediction on the test set, predicted intents were compared to the actual intents to determine the number of correct responses. The accuracy of the English model was 80%, while the accuracy of the Arabic model was 75%.



(a)

(b)

Fig. 3. Figure 1 Screenshot of an English and Arabic Conversation.

TABLE IV. SAMPLE QUERIES FROM THE TEST SET

Query	actual_intent (label)
If I'm working do I still need to take work placement course?	wp-working-en
how long is a summer semester?	sum-duration-en
امت اضافة حذف	add-drop-ar
كيف اعرف اذا الدورة هيبرد او وجها لوجه	course-how-ar

In the second evaluation method, we involved the end-users, students, and advisors, to test the chatbot GUI application and performance of the prediction model. The objective of this evaluation was threefold, to test the chatbot interface, the effectiveness of the conversational system, and the accuracy of responses based on context.

Thirty students and three advisors evaluated the chatbot both in English and Arabic language. The participants were briefed on the context of the chatbot corpus and asked to provide random queries. The interactions were recorded in a CSV file along with the response's predicted intent, context, and confidence. When the response confidence was below 0.75, the chatbot requested the user to rephrase the question if they thought the response was not accurate. In nearly 20% of the cases, the chatbot engine could not determine the intent due to out of vocabulary words or out-of-context queries, so the standard response “Contact your advisor” was displayed. This result shows that it is essential for the chatbot corpus to be extended to include a wider domain of queries. From all the captured test inputs, we considered only the intents within the context specified to determine the accuracy of the response.

Our study does not evaluate the user satisfaction of the chatbot application. This type of study involves gathering empirical feedback from end-users from the Human-Computer Interaction perspective, which is outside the scope of our paper. However, during the testing phase of the chatbot application, several students commented that they found the chatbot useful and would prefer to use it instead of going to their advisor. In addition, they appreciated the quick response and constant availability of a chatbot application. Another observation we made from this evaluation is that students preferred to use English rather than Arabic when writing their queries as it was faster for them to type. There were several words that they did not know how to write in Arabic, such as "probation” or “covid.”.

V. CONCLUSION

In today's world, conversational agents are proving to be one of the most innovative forms of user interaction. This paper presents a new bilingual task-oriented, domain-specific Arabic English chatbot explicitly designed to advise university students to ease their academic journey. The chatbot uses NLP and neural network algorithms to retrieve English or Arabic responses. Through the bot, students may communicate and receive responses to their inquiries. Two chatbot models have been created in Python using a supervised deep learning algorithm, trained on English and Arabic corpora, respectively. An Arabic and English corpus of roughly 152 intents in both English and Arabic has been developed, with 356 patterns. In order to train the model, we pre-processed the training data and encoded each intent using the Python library so that it is suitable for the neural network algorithm. In the absence of a good library in Python for lemmatizing Arabic words, ISRIStemmer was used to stem the words. We use three logical components (NLP), (NLU), and (NLG) in our chatbot engine in order to pre-process the input query and to predict and generate a response based on the user's request. The prediction error threshold was set at 25%, and all predictions with probabilities above this threshold were accepted.

Moreover, the chatbots graphical user interface was developed using the Python tkinter library to interact with the user and display the most appropriate response. Two types of evaluations were performed to measure the performance of the system; the confidence score and another automated evaluation performed by the system users. The first provides 80% accuracy in English and 75% in Arabic. The second evaluation performed by the user also has similar results.

A. Limitations and Future Work

The bilingual chatbot system has some limitations. It was challenging to spellcheck Arabic, and many errors were produced when the results did not match the input inquiry after the check was performed. There was another issue with lemmatizing in Arabic. Some of the words did not retain their meaning, so the response was incorrect. There were also challenges with getting a response when the model confidence level was low, and the model did not understand the user's intent.

Both Arabic and English corpora should be expanded to include more vocabulary in each intent tag. Additionally, adding more intents with new context will broaden the scope of the corpora used in English and Arabic and expand advisory areas. Finally, the Arabic spellchecker needs further study and analysis to be used in the system.

Another limitation of the study is that the developed chatbot does not provide personalized assistance to students. Future work would enhance the chatbot with intelligent capabilities that allow personalized responses containing information such as students' GPA, academic standing, and courses required for graduation. Such a chatbot could assist advisors in developing study plans and communicating with the students. Another enhancement to the chatbot that can add value to the communication is to send push notifications to remind students of upcoming deadlines for registration, add and drop periods, and more.

REFERENCES

[1] M. Allouch, A. Azaria, and R. Azoulay, “Conversational Agents: Goals Technologies and Challenges,” Sensors (Switzerland), pp. 1–48, 2021.

[2] M. Adam, M. Wessel, and A. Benlian, “AI-based chatbots in customer service and their effects on user compliance,” Electron. Mark., vol. 31, no. 2, pp. 427–445, 2021, doi: 10.1007/s12525-020-00414-7.

[3] J. Grudin and R. Jacques, “Chatbots, humbots, and the quest for artificial general intelligence,” Conf. Hum. Factors Comput. Syst. - Proc., pp. 1–11, 2019, doi: 10.1145/3290605.3300439.

[4] S. Campbell and C. Nutt, “Academic Advising in the New Global Century: Supporting Student Engagement and Learning Outcomes Achievement,” Peer Rev., vol. 10, no. 1, p. 4, 2008.

[5] J. K. Drake, “The Role of Academic Advising in Student Retention and Persistence,” About Campus Enrich. Student Learn. Exp., vol. 16, no. 3, pp. 8–12, 2011, doi: 10.1002/abc.20062.

[6] T. Fricker, “The Relationship between Academic Advising and Student Success in Canadian Colleges: A Review of the Literature.,” Coll. Q., vol. 18, no. 4, p. n4, 2015.

[7] O. Iatrellis, A. Kameas, and P. Fitsilis, “Academic advising systems: A systematic literature review of empirical evidence,” Educ. Sci., vol. 7, no. 4, 2017, doi: 10.3390/educsci7040090.

[8] A. Miklosik, N. Evans, A. Mahmood, and A. Qureshi, “The Use of Chatbots in Digital Business Transformation : A Systematic Literature Review,” IEEE Access, vol. 9, pp. 106530–106539, 2021, doi: 10.1109/ACCESS.2021.3100885.

[9] M. R. Pacheco-Lorenzo, S. M. Valladares-Rodríguez, L. E. Anido-Rifón, and M. J. Fernández-Iglesias, “Smart conversational agents for the detection of neuropsychiatric disorders: A systematic review,” J. Biomed. Inform., vol. 113, p. 103632, 2021, doi: https://doi.org/10.1016/j.jbi.2020.103632.

[10] M. Jovanovic, M. Baez, and F. Casati, “Chatbots as Conversational Healthcare Services,” IEEE Internet Comput., vol. 25, no. 3, pp. 44–51, 2021, doi: 10.1109/MIC.2020.3037151.

[11] C. W. Okonkwo and A. Ade-Ibijola, “Chatbots applications in education: A systematic review,” Comput. Educ. Artif. Intell., vol. 2, p. 100033, 2021, doi: 10.1016/j.caeai.2021.100033.

[12] R. S. Wallace, “The anatomy of ALICE,” in Parsing the turing test, Springer, 2009, pp. 181–210.

[13] M. Daswani, K. Desai, M. Patel, R. Vani, and M. Eirinaki, “CollegeBot: A Conversational AI Approach to Help Students Navigate College,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12424 LNCS, no. October 2020, pp. 44–63, 2020, doi: 10.1007/978-3-030-60117-1_4.

[14] N. Asghar, P. Poupart, J. Hoey, X. Jiang, and L. Mou, “Affective neural response generation,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10772 LNCS, pp. 154–166, 2018, doi: 10.1007/978-3-319-76941-7_12.

[15] X. Zhou and W. Y. Wang, “MOJITALK: Generating Emotional Responses at Scale,” Proc. 56th Annu. Meet. Assoc. Comput. Linguist., pp. 1–2, 2018.

[16] S. Hussain, O. Ameri Sianaki, and N. Ababneh, A Survey on Conversational Agents/Chatbots Classification and Design Techniques, vol. 927, no. March. Springer International Publishing, 2019.

[17] E. Adamopoulou and L. Moussiades, “An Overview of Chatbot Technology BT - Artificial Intelligence Applications and Innovations,” 2020, pp. 373–383.

[18] E. M. Latorre-Navarro and J. G. Harris, “An Intelligent Natural Language Conversational Agent System for Academic Advising,” Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 1, 2015.

[19] S. Alias, M. S. Sainin, T. S. Fun, and N. Daut, “Intent Pattern Discovery for Academic Chatbot-A Comparison between N-gram model and Frequent Pattern-Growth method,” ICETAS 2019 - 2019 6th IEEE Int. Conf. Eng. Technol. Appl. Sci., 2019, doi: 10.1109/ICETAS48360.2019.9117315.

[20] B. R. Ranoliya, N. Raghuwanshi, and S. Singh, “Chatbot for university related FAQs,” 2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017, vol. 2017-Janua, pp. 1525–1530, 2017, doi: 10.1109/ICACCI.2017.8126057.

[21] C. Asakiewicz, E. A. Stohr, and S. Mahajan, “Building a Cognitive Application Using Watson DeepQA,” IT Prof., vol. 19, no. 4, pp. 36–44, 2017.

[22] S. Meshram, N. Naik, V. R. Megha, T. More, and S. Kharche, “College Enquiry Chatbot using Rasa Framework,” in IEEE Asian Conference on Innovation in Technology (ASIANCON), 2021, pp. 1–8, doi: 10.1109/ASIANCON51346.2021.9544650.

[23] D. A. Ali and N. Habash, “Botta: An Arabic dialect chatbot,” COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Syst. Demonstr., pp. 208–212, 2016.

[24] S. AlHumoud, A. Al Wazrah, and W. Aldamegh, “Arabic Chatbots: A survey,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 8, pp. 535–541, 2018, doi: 10.14569/ijacsa.2018.090867.

[25] D. Al-Ghadhban and N. Al-Twairesh, “Nabiha: An Arabic dialect chatbot,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 3, pp. 452–459, 2020, doi: 10.14569/ijacsa.2020.0110357.

[26] W. El Hefny, Y. Mansy, M. Abdallah, and S. Abdennadher, “Jooka: A Bilingual Chatbot for University Admission,” World Conf. Inf. Syst. Technol., vol. 1367 AISC, no. March, pp. 671–681, 2021, doi: 10.1007/978-3-030-72660-7_64.

[27] K. Ralston, Y. Chen, H. Isah, and F. Zulkernine, “A voice interactive multilingual student support system using IBM watson,” in 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019, 2019, pp. 1924–1929, doi: 10.1109/ICMLA.2019.00309.

[28] S. Badlani, T. Aditya, M. Dave, and S. Chaudhari, “Multilingual healthcare chatbot using machine learning,” 2021 2nd Int. Conf. Emerg. Technol. INCET 2021, pp. 1–6, 2021, doi: 10.1109/INCET51464.2021.9456304.

[29] J. K. Catapang, G. A. Solano, and N. Oco, “A Bilingual Chatbot Using Support Vector Classifier on an Automatic Corpus Engine Dataset,” 2020 Int. Conf. Artif. Intell. Inf. Commun. ICAIIC 2020, pp. 187–192, 2020, doi: 10.1109/ICAIIC48513.2020.9065208.

[30] A. R. Martinez, “A Survey of the Usages of Deep Learning for Natural Language Processing,” IEEE Trans. Neural Networks Learn. Syst., vol. 32, no. 3, pp. 352–357, 2021, doi: 10.1002/wics.76.

[31] L. Bottou, “18 Stochastic Gradient Descent Tricks,” pp. 421–422, 2012.

[32] Y. Wang, Y. Li, Y. Song, and X. Rong, “The influence of the activation function in a convolution neural network model of facial expression recognition,” Applied Sciences (Switzerland), vol. 10, no. 5. 2020, doi: 10.3390/app10051897.

[33] M. D. Zeiler, “ADADELTA: An Adaptive Learning Rate Method,” 2012, [Online]. Available: http://arxiv.org/abs/1212.5701.

[34] P. Huo, Y. Yang, J. Zhou, C. Chen, and L. He, “TERG: Topic-Aware Emotional Response Generation for Chatbot,” 2020, doi: 10.1109/IJCNN48605.2020.9206719.

[35] P. Huo, Y. Yang, J. Zhou, C. Chen, and L. He, “TERG: Topic-Aware Emotional Response Generation for Chatbot,” 2020, doi: 10.1109/IJCNN48605.2020.9206719.

Modified Prophet+Optuna Prediction Method for Sales Estimations

Kohei Arai[1], Ikuya Fujikawa[2], Yusuke Nakagawa[3], Tatsuya Momozaki[4], Sayuri Ogawa[5]
Information Science Department, Saga University, Saga City, Japan[1]
Success Institute Chain: SIC Co., Ltd, Hakata-ku, Fukuoka City, Fukuoka, Japan[2, 3, 4, 5]

***Abstract*—A prediction method for estimation of sales based on Prophet with a consideration of nonlinear events and conditions by a modified Optuna is proposed. Linear prediction does not work for a long-term sales prediction because purchasing actions are based on essentially nonlinear customers' behavior. One of nonlinear prediction methods is the well-known Prophet. It, however, is still difficult to adjust the nonlinear parameters in the Prophet. To adjust the parameters, the Optuna is widely used. It, however, is not good enough for parameter tuning by the Optuna. Therefore, the Optuna is modified with a short-term moving mean and standard deviation of the sales for final prediction. More than that, specific event such as typhoon event is to be considered in the sales prediction. Through experiments with a real sales data, it is found the sensitivity of the parameters the upper window, lower window, event dates, etc. for the final sales and the effect of the Optuna is 11.73%. Also, it is found that the effect of the consideration of Covid-19 is about 2.4% meanwhile the effect of the proposed modified Optuna is around 3 % improvement of the prediction accuracy (from 80 % to 83 %).**

Keywords*—Prediction method; nonlinearity; prophet; optuna; typhoon event; modified optuna; mean and standard deviation adjustment

I. Introduction

Periodicity, event effects, long-term trends, and outliers are not limited to this data, but are common features of general time series data. When creating a model for future prediction, it is necessary to incorporate these features into the model well. Prophet models each of the four features and combines them to predict future values. Such a model is called a Generalized Additive Model.

Prophet is a library for time series analysis developed by Facebook's Core Data Science team in 2017[1]. Libraries are provided in both Python and R. In addition, this Prophet is embedded as a template in AutoML services such as AWS, Azure, and DataRobot for flexible modeling in future forecasting tasks[2].

There are five advantages of Prophet:

1) It can be made a model without knowledge of statistics: Simply specify the data and perform the training to complete the model.

2) Easy to incorporate domain knowledge: It can be easily put in the domain knowledge that the data analyst has.

3) No feature engineering required: Prophet training uses minimally preprocessed data. There is no need to remove trend components or convert to a moving average series.

4) There is no problem even if there are missing values: Even if there is a defect in the training data, no error will occur, and training will be performed normally. Therefore, it is not necessary to fill in the missing values in advance.

5) Easy to interpret prediction results: Prophet is a model that adds four terms. Each term represents a trend, periodicity, event effect, and error, and after prediction, the components can be extracted for each term and the obtained prediction results can be considered.

On the other hand, Optuna is a Bayesian optimization package created for optimizing hyperparameters of machine learning models[3]. It performs optimization using TPE, which is a new method among Bayesian optimizations. It can be easily used in a single process, or it can be learned in parallel on many machines. When performing parallel processing, this is achieved by creating an Optuna file on the database and referencing it from multiple machines, so it is wonderful that all machines that can access the DB can participate in learning[4].

The proposed nonlinear prediction method is based on Prophet with Optuna for parameter tuning. It is not easy to optimize the parameters in Prophet and is not ensure the best fit parameters for Prophet. In this paper, therefore, some programmatic method for the parameter optimization is proposed and effectiveness of the proposed method is validated with a nonlinear sales data.

The biggest challenge of this research work is to predict one year term of sales (annual amount of sales). Although there are many prediction methods which allow to predict one day after the current time, there is no such method which allows forecast ahead for the following 365 days with an acceptable prediction accuracy. Therefore, nonlinearity, seasonal effect, event effect, the other influencing factors have to be considered.

One of nonlinear prediction methods is the well-known Prophet. It, however, is still difficult to adjust the nonlinear parameters in the Prophet. To adjust the parameters, the Optuna is widely used. It, however, is not good enough for

[1]https://facebook.github.io/prophet/
[2]https://peerj.com/preprints/3190/
[3] https://optuna.readthedocs.io/en/stable/tutorial/first.html
[4] https://optuna.readthedocs.io/en/stable/reference/index.html

parameter tuning by the Optuna. Therefore, the Optuna is modified with a short-term moving mean and standard deviation of the sales for final prediction.

In the next section, related research works are reviewed followed by the proposed method. Then, the fact that a linear prediction method does not work for nonlinear time series of data is shown. After that, the validation of effectiveness of the proposed method is described together with effectiveness of the Optuna. Finally, conclusion and some discussions are described followed by future research work.

II. Related Research Work

There are the following related research works on prediction,

Probabilistic cellular automata-based approach for prediction of hot mudflow disaster area and volume is proposed [1]. New approach of prediction of Sidoarjo hot mudflow disaster area based on probabilistic cellular automata is also proposed [2]. On the other hand, GIS based 2D cellular automata approach for prediction of forest fire spreading is proposed [3].

Cell based GIS as cellular automata for disaster spreading prediction and required data systems is investigated [4] together with hot mudflow prediction area model and simulation based cellular automata for LUSI and plume at Sidoarjo East Jawa [5].

Comparative study between eigen space and real space-based image prediction methods by means of autoregressive model is conducted [6] together with comparative study on image prediction methods between the proposed morphing utilized method and Kalman Filtering method [7].

Prediction method for time series of imagery data in eigen space is proposed [8]. Meanwhile, image prediction method with non-linear control lines derived from Kriging method with extracted feature points based on morphing is proposed [9]. On the other hand, cell-based GIS as cellular automata for disaster spreading predictions and required data systems is proposed [10].

Prediction method of El Nino Southern Oscillation event by means of wavelet-based data compression with appropriate support length of base function is proposed [11]. On the other hand, Question Answering for collaborative learning with answer quality prediction is created [12].

Wildlife damage estimated and prediction method using blog and tweet information is proposed [13]. Prediction method for large diatom appearance with meteorological data and MODIS derived turbidity as well as chlorophyll-a in Ariake Bay area in Japan is proposed and validated [14].

Method for thermal pain level prediction with eye motion using SVM is proposed [15] together with prediction method for large diatom appearance with meteorological data and MODIS derived turbidity and chlorophyll-a in Ariake bay area in Japan [16].

Smartphone image based agricultural product quality and harvest amount prediction method is proposed [17]. Meanwhile, data retrieval method based on physical meaning and its application for prediction of linear precipitation zone with remote sensing satellite data and open data is also proposed [18].

Recursive Least Square: RLS method-based time series data prediction for many missing data is proposed [19]. Furthermore, prediction of isoflavone content in beans with Sentinel-2 optical sensor data by means of regressive analysis is proposed and conducted [20].

III. Proposed Prediction Method based on Prophet

A library often used for time series analysis using AI, especially for future prediction with the following features,

1) There is periodicity, weekly and yearly periodicity
2) There is an event effect
3) There is a long-term trend
4) There are outliers (noise)

Prophet is developed by Facebook in 2017.

The model formula of Prophet is as follows,

$$y=g(t)+s(t)+h(t)+\varepsilon_t \quad (1)$$

where *y(t)*: variable for prediction, *g(t)*: trend, *s(t)*: periodic, *h(t)*: event effect, ε_t: normal distribution of noise.

Basically, it can be used as the same as scikit-learn. Model instance creation can be done with fit flow then creation is also done with a data frame for prediction. After that, a prediction is made with the predict method.

Nonlinear for trend term: Linear by default Specify the upper limit of prediction (cap) to make it nonlinear. Fig. 1 shows the illustrative view of determination of upper limit of prediction.

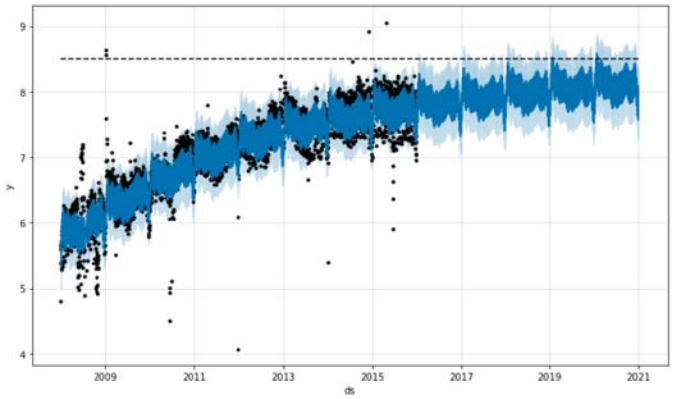
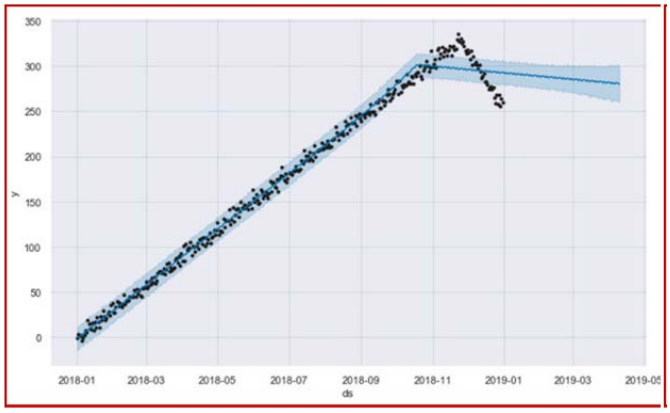


Fig. 1. Illustrative View of Determination of Upper Limit of Prediction.

1) Changepoint-range: The default setting does not reflect the most recent change point as shown in Fig. 2 (a). The data used by Prophet to estimate the trend change point is 80% of the total by default. Resolved by setting "changepoint-range = 1". It tends to be predicted that the latest data will be overwhelmed as shown in Fig. 2 (b).



(a)Default

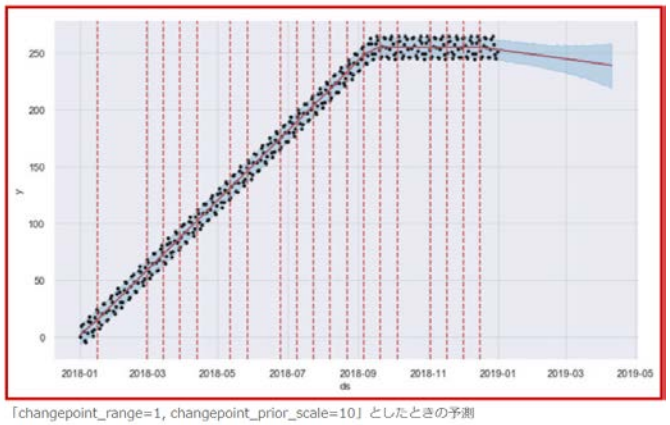
(b) Changepoint-range = 1

Fig. 2. Changepoint-Range Setting.

2) Changepoint-prior-scale: It represents the variance of the Laplace distribution, which is the prior distribution of the trend term, and the larger it is, the easier it is to detect the change point as shown in Fig. 3 (a).,

3) n-changepoints: It represents the number of change point candidates to be detected, and the larger the number, the easier it is to detect more change points as shown in Fig. 3 (b).

(a) Changepoint-Prior Scale.



(b) n-changepoints

Fig. 3. Another Parameter Setting.

Trend term g(t) is represented as equation (2) and can be determined as follows,

$$g(t) = \frac{C}{1+\exp(-k(t-m))} \quad (2)$$

where *C*: Upper limit, *k:* Growing ratio, *m*: Offset

This is the base logistic curve which is shown in Fig. 4 Phenomenon with a flow of less at the beginning then more in the middle, less again after that.

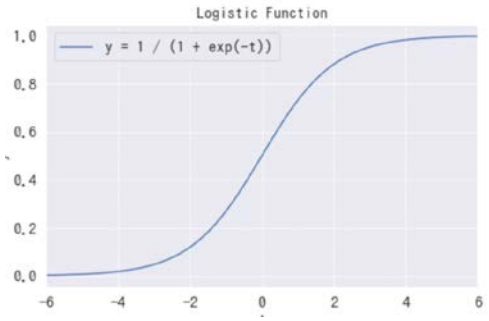


Fig. 4. Logistic Function.

Upper limit, growing ratio and offset are determined as follows,

Since *m* is an expression that directly subtracts the value of *t* as shown in Fig. 5 (c). The curve simply moves from side to side.

There are seasonal fluctuations. There is periodicity. It can be expressed like signal processing.

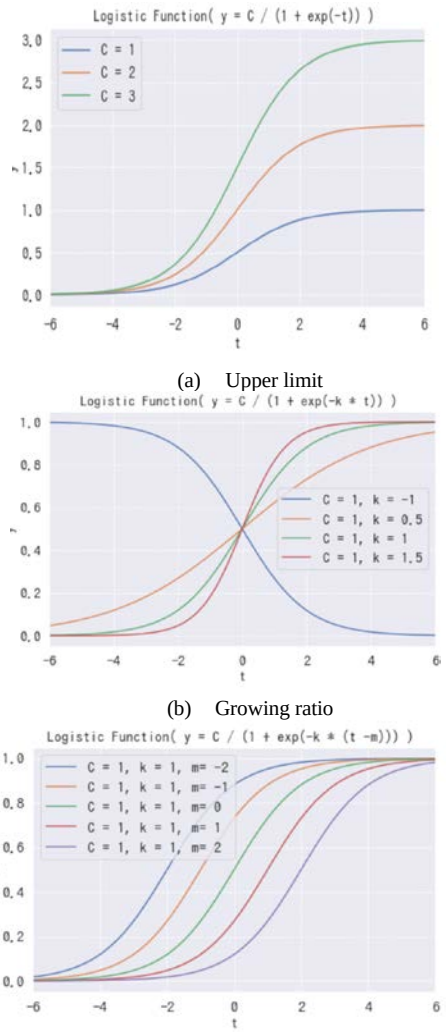
$$s(t) = \sum_{n=1}^{N}(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin(\frac{2\pi nt}{P})) \quad (3)$$

Fit with N = 10 for a yearly cycle and N = 3 for a weekly cycle.

$$\beta = (a_1, b_1, \dots, a_N, b_N)^T \quad (4)$$

$$X(t) = [\cos\left(\frac{2\pi(1)t}{365.25}\right), \sin\left(\frac{2\pi(1)t}{365.25}\right), \dots, cos\left(\frac{2\pi(10)t}{365.25}\right), sin\left(\frac{2\pi(10)t}{365.25}\right)] \quad (5)$$

$$s(t) = X(t)\beta \quad (6)$$



(a) Upper limit

(b) Growing ratio

(c) Offset

Fig. 5. Determination of Upper Limit, Growing Ratio and Offset Determinations.

Event effect is defined by incorporate sudden event effects into the model *h(t)*. Prophet is designed so that the analyst can create a list of event calendars and incorporate them into the model. The coefficient parameter for each event *i* is κ_i, and the vector is represented by κ.

$$D_i = (\dots, 1975/12/25, 1976/12/25, \dots, 2020/12/25, \dots) \quad (7)$$

$$Z(t) = [1(t \in D_1), \dots, 1(t \in D_L)] \quad (8)$$

$$h(t) = Z(t)\kappa, \kappa \sim Normal(0, v^2) \quad (9)$$

Probability Density Function: PDF is defined as follows,

If you assume the distribution for each parameter, you can treat it as a state space model. In fact, in Prophet, this model formula is described in Stan and optimized by the L-BFGS method etc.

$$\beta \sim Normal(0, \sigma^2) \quad (10)$$

IV. Experiments

A. Example of Sales Prediction and Sensitivity of the Parameters

Through an adjustment of the Prophet parameters to forecast Mega-Donki Hair Salon (One of the Hair Salons in concern) sales, and then prediction of the sales. Fig. 6 shows the prediction result. In this case, the sales data of 2015 to 2020 is used for training data and also the sales data of 2021 is used for validation data.

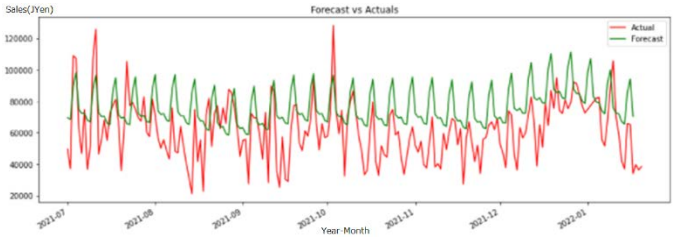


Fig. 6. Prediction of Mega-Donki Hair Salon Sales in 2021.

There is a systematic error. Also, prediction error is not so small. Therefore, some parameter adjustments are required to improve the prediction accuracy.

Optuna is a software framework for automating hyperparameter optimization. The author adjusted the parameters of Prophet using Optuna and tried to forecast the sales of the Mega-Donki Hair Salon. Parameters that can be tuned such as seasonal prior distribution, degree of influence, range of use of data used for detection of change points, influence of trends, etc. After the adjustment by Optuna, prediction result is improved as shown in Fig. 7. The total prediction error is reduced from 38.09 to 26.36. Namely, 11.73 % of improvement is confirmed.

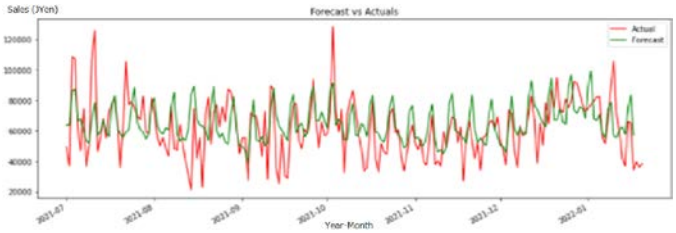


Fig. 7. Prediction of Mega-Donki Hair Salon Sales after the Parameter Adjustment with Optuna.

Tuned parameters are as follows,

1) Changepoint-range: Percentage of what range of data is used to detect the trend change point. The default is 0.8, which uses the first 80% of the data to detect the trend change point. According to the formula [0.8, 0.95] Range is reasonable

2) n-changepoints: A parameter that represents the number of candidates change points to detect. The larger the parameter, the easier it is to detect more change points. The default is 25.

3) Changepoint-prior-scale: Parameters that control the flexibility of the trend. If it is too small, the trend will be inadequate, and if it is too large, the trend will be overfitted. The default is 0.05, and the formula says that the range [0.001, 0.5] is reasonable.

4) Seasonality-prior-scale: Parameters that control seasonal flexibility. The default is 10, and the formula says that the range [0.01, 10] is reasonable.

Add-seasonality (period, Fourier-order): Not only the specified year / week / day periodicity, but also the model of any cycle can be set by the user. For each periodicity, the unit of periodicity (period), the Fourier series that is the basis of the seasonal component (Fourier-order), and the degree of influence of seasonality (prior-scale) are set. In order to be able to adjust the Fourier series and the degree of influence of each seasonality, all the prescribed seasonality is set to False, and weekly, monthly, yearly, and quarterly periodic fluctuations are added.

B. Sales Prediction Results with Parameter Setting

Data preprocessing (missing value completion) is needed. Then, consideration of event effect (entrance ceremony, graduation ceremony, pension payment date) is necessary. There is a date with zero sales. Until now, it was excluded and calculated. Therefore, complement the interpolation by the average value is needed. The average value is the day of zero. Complemented with the average value of the day of the week. With lower-window and upper-window, the range can be extended the range to which the event effect is applied to the days around the event day. Also, if Christmas is set as an event and lower window is set to -1, the event effect can be applied until Christmas Eve.

The date of the pension payment can be considered. After the 15th of even-numbered months. Also, one week defined as (7 days). Improvement of the prediction accuracy (MAPE: Mean Absolute Prediction Error) for the specific two shops: Konoha Mall Hashimoto Hair Salon (This Hair Salon is another Hair Salon in concern and has low prediction accuracy).

1) Before considering the event: MAPE= 37.8

2) After considering the entrance ceremony, graduation ceremony, and pension: MAPE=36.4

For the Hakata Station South Hair Salon (Another Hair Salon in concern) case,

1) Before considering the event: MAPE=25.5

2) After considering the entrance ceremony, graduation ceremony, and pension: MAPE=24.7

3) After considering the entrance ceremony and pension, MAPE=24.8

Sales are declining near the graduation ceremony due to learning from training data. The sales are increasing near the entrance ceremony. Also, sales increase (upside) can be dealt with relatively, but sales decrease (downside) cannot be dealt with (red ... actual green ... forecast) Hakata-eki-minami Hair Salon (another Hair Salon in concern) data as shown in Fig. 8. There are some strange prediction errors marked with blue ellipsoids. For instance, sales are gotten down on August 25 due to the typhoon #15 is hit over these areas as shown in Fig. 9. These events can be considered in the prediction by Prophet.

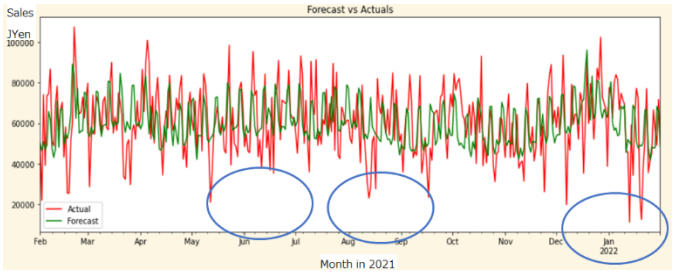


Fig. 8. Sales Prediction Result for Hakata-eki-minami Hair Salon.

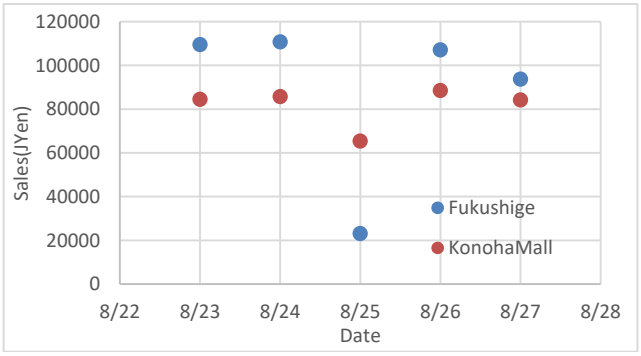
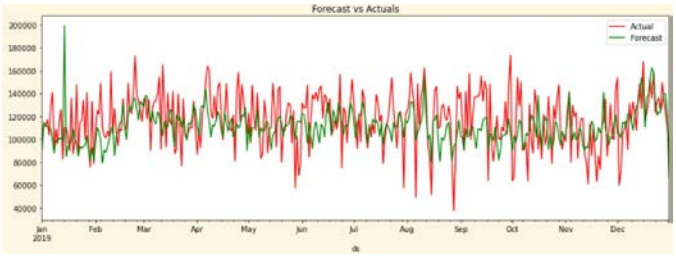


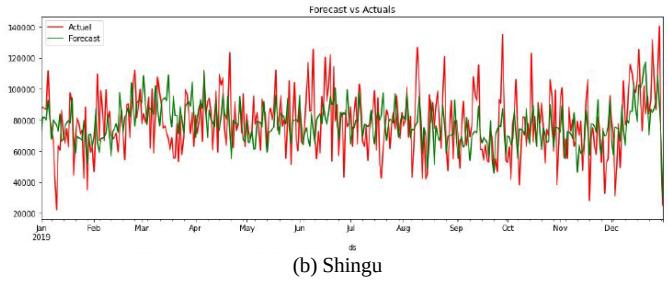
Fig. 9. Sales of the Fukushige and the Konoha-Mall Hair Salons during from Aug.22 to Aug. 28.

C. Influence Due to the Covid-19

The sales of the hair salon have been changed due to the Covid-19. To investigate the influence of Covid-19, the sales of the Fukushige hair salon have been predicted for one year of 2019 utilizing the nine years sales data, 2010 to 2018. The actual and the predicted sales with Prophet and the proposed modified Optuna are shown in Fig. 10(a). As the result, it is found that the MAPE is improved from 18.09 to 16.61. Also, as shown in Fig. 10(b), it is found that the actual and the predicted sales of Shingu hair salon is improved from 23.6 to 21.1.



(a) Fukusige



(b) Shingu

Fig. 10. Influence of Covid-19 on the Sales Prediction.

V. Conclusion

Through the experiments, it is found the followings,

1) Although Prophet is basically linear prediction method, it does work because it can consider trend, seasonal changes, event effect.

2) Upper window, lower window, event dates, etc. need to be entered from specialized knowledge and experience.

3) The proposed Optuna parameter tuning shows 11.73% of improvement in mean prediction error for the specific Hair Salon sales in comparison to the Prophet prediction without Optuna.

4) When the events of typhoons, heavy rain, pension payment date, entrance ceremonial date, etc. are considered in the proposed Optuna parameter tuning, then the sale prediction error is reduced.

5) The effect of the considering the entrance ceremony, graduation ceremony, and pension payment days is less than 2%.

6) Influence of Covid-19 on the sales prediction is clarified. If the influence is considered in the sales prediction processes, MAPE is improved from 8.2 to 10.6 %.

VI. Future Reaesrch Work

Further investigations are required for improvement of prediction accuracy by considering the other influencing factors such as coupon, special campaign, etc. to the sales. Weather forecast data, geolocation, population, environmental factors are other candidates of the influencing factors.

Acknowledgment

The authors would like to thank to the cooperative staff of the Success Institute Chain: SIC Holdings Co., Ltd, for their effort to conduct this research work. Also, the authors would like to thank to Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

References

[1] Achmad Basuki, Tri Harsono and Kohei Arai, Probabilistic cellular automata based approach for prediction of hot mudflow disaster area and volume, Journal of EMITTER1, 1, 11-20, 2010.

[2] Kohei Arai, Achmad Basuki, New Approach of Prediction of Sidoarjo Hot Mudflow Disaster Area Based on Probabilistic Cellular Automata, Geoinformatica - An International Journal (GIIJ), 1, 1, 1-11, 2011.

[3] Kohei Arai, Achmad Basuki, GIS based 2D cellular automata approach for prediction of forest fire spreading, International Journal of Research and Reviews on Computer Science, 2, 6, 1305-1312, 2011.

[4] Kohei Arai, Cell based GIS as Cellular Automata for disaster spreading prediction and required data systems, CODATA Data Science Journal, 137-141, 2012.

[5] Kohei Arai, A.Basuki, T.Harsono, Hot mudflow prediction area model and simulation based cellular automata for LUSI and plume at Sidoarjo East Jawa, Journal of Computational Science (Elsevior) 3,3,150-158, 2012.

[6] Kohei Arai, Comparative Study between Eigen Space and Real Space Based Image Prediction Methods by Means of Autoregressive Model, International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 3, No. 6, 1869-1874, December 2012, ISSN: 2079-2557.

[7] Kohei Arai, Comparative Study on Image Prediction Methods between the Proposed Morphing Utilized Method and Kalman Filtering Method, International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 3, No. 6, 1875-1880, December 2012, ISSN: 2079-2557.

[8] Kohei Arai Prediction method for time series of imagery data in eigen space, International Journal of Advanced Research in Artificial Intelligence, 2, 1, 12-19, (2013).

[9] Kohei Arai Image prediction method with non-linear control lines derived from Kriging method with extracted feature points based on morphing, International Journal of Advanced Research in Artificial Intelligence, 2, 1, 20-24, (2013).

[10] Kohei Arai, Cell based GIS as cellular automata for disaster spreading predictions and required data systems, Advanced Publication, Data Science Journal, Vol.12, WDS 154-158, 2013.

[11] Kohei Arai, Prediction method of El Nino Southern Oscillation event by means of wavelet based data compression with appropriate support length of base function, International Journal of Advanced Research in Artificial Intelligence, 2, 8, 16-20, 2013.

[12] Kohei Arai, Anik Nur Handayani, Question Answering for collaborative learning with answer quality prediction, International Journal of Modern Education and Computer Science, 5, 5, 12-17, 2013.

[13] Kohei Arai, Shohei Fujise, Wildlife Damage Estimated and Prediction Using Blog and Tweet Information, International Journal of Advanced Research on Artificial Intelligence, 5, 4, 15-21, 2016.

[14] Kohei Arai, Prediction method for large diatom appearance with meteorological data and MODIS derived turbidity as well as chlorophyll-a in Ariake Bay area in Japan, International Journal of Advanced Computer Science and Applications IJACSA,8,3,39-44, 2017.

[15] Kohei Arai, Method for Thermal Pain Level Prediction with Eye Motion using SVM, International Journal of Advanced Computer Science and Applications IJACSA, 9, 4, 170-175, 2018.

[16] Kohei Arai, Prediction method for large diatom appearance with meteorological data and MODIS derived turbidity and chlorophyll-a in Ariake bay area in Japan, International Journal of Advanced Computer Science and Applications IJACSA, 10, 9, 39-44, 2019.

[17] Kohei Arai, Osamu Shigetomi, Yuko Miura, Satoshi Yatsuda, Smartphone image based agricultural product quality and harvest amount prediction method, International Journal of Advanced Computer Science and Applications IJACSA, 10, 9, 24-29, 2019.

[18] Kohei Arai, Data Retrieval Method based on Physical Meaning and its Application for Prediction of Linear Precipitation Zone with Remote Sensing Satellite Data and Open Data, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 10, 56-65, 2020.

[19] Kohei Arai, Kaname Seto, Recursive Least Square: RLS Method-Based Time Series Data Prediction for Many Missing Data, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 11, 66-72, 2020.

[20] Kohei Arai, Prediction of Isoflavone Content in beans with Sentinel-2 Optical Sensor Data by Means of Regressive Analysis, Proceedings of SAI Intelligent Systems Conference, IntelliSys 2021: Intelligent Systems and Applications pp 856-865, 2021.

Authors' Profile

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 77 books and published 680 journal papers as well as 550 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA http://teagis.ip.is.saga-u.ac.jp/index.ht

Encryption Algorithms Modeling in Detecting Man in the Middle Attack in Medical Organizations

Sulaiman Alnasser, Raed Alsaqour
Department of Information Technology, College of Computing and Informatics, Saudi Electronic University, Riyadh 93499, Saudi Arabia

***Abstract*—A cyberattack is a serious crime that could affect medical organizations. These attacks could affect medical organization sensitive data disclosure, loss of organization data, or the business's continuity. The Man-in-The-Middle (MITM) attack is one of the threats that could impact medical organizations. It happens when unapproved outsiders break into the traffic between two parties that think they are conversing directly. At the same time, the adversary can access, read, and change secret information. Because of that, medical organizations lose confidentiality, integrity, and availability. Data encryption is a solution that changes vital information to unreadable by unauthorized and unintended parties. It could involve protecting data with cryptography, usually by leveraging a scrambled code. Only the individuals with the decoding key can read the information. There is no full protection due to the variety of MITM attacks. Each encryption algorithm has its advantages and disadvantages, like the speed of encryption and decryption, strength of the algorithm, and the cipher type. This research investigates the MITM attacks and comprehensively compares the Rivest Shamir Adleman algorithm and the Elliptic Curve Cryptography algorithm.**

Keywords*—*Cyberattack; medical organization; man in the middle attack; encryption algorithm; rivest shamir adleman algorithm; elliptic curve cryptography algorithm

I. Introduction

In the contemporary organizational environment, governments, businesses, medical, and individuals all store data in electronic form. Electronic data storage is more effective than the previous physical storage forms since it is more compact, allows instantaneous transfer, and is easier to access information via databases [1]. Over time, the value of data increases, and organizations and individuals widely recognize stored data as among the most valuable items that must be protected against all potential threats. However, with such a notable electronic revolution, effective data storage and management face multiple new security threats that are potentially more damaging [2]. For example, electronic data has a high risk of being copied, leaving the original unaltered, or stolen, and has a high vulnerability for interceptions and alterations. Therefore, an effective data security measure must enhance secrecy, integrity, and availability. Part of the technical services crucial for optimizing data protection include data authentication and encryption [3].

A Man-in-The-Middle (MITM) attack is one of the threats that could impact medical organizations [4]. It happens when unapproved outsiders break into the traffic between two parties that think they are conversing directly. At the same time, the adversary can access, read, and change secret information. Because of that, organizations lose confidentiality, integrity, and availability. Data encryption is a powerful solution to eliminate the MITM attack [5]. It encompasses translating or encoding data into another form or a code to ensure that it is only accessible to persons with access to the secret key. It is a robust approach to protecting private information and sensitive data. It enhances communication security between different parties and servers. Encrypted data is largely depicted as ciphertext. The process is the most effective and popular information security method [6].

Data encryption is central in enhancing and maintaining the confidentiality of sensitive and private information, and the technology also increases data safety among remote workers. Therefore, it is an essential security safeguard for corporations, and in the long term, it positively impacts consumer trust and overall profitability [7].

Asymmetric encryption algorithms can be complicated, especially; most businesses and individuals rely on this type of encryption since they are strong and hard to break. Unfortunately, studies reveal that in the contemporary technology-infiltrated market setting, a wide range of cybersecurity issues and threats negatively affect entities' effective functionality [7]. However, with data encryption being done properly, for instance, by leveraging a high enough level of encryption and adequate safeguarding of the respective encryption key, the security and privacy of various features can be enhanced. This is vital in eliminating potential threats that could ultimately compromise data safety and security. Furthermore, the Internet of things (IoTs) started to become a valid solution in the medical field. Currently, many surgeries are done remotely [8]. So, strong, accurate, and speed algorithms are primary conditions that medical sectors cannot abandon.

This study presents a critical analysis of encryption. It provides cases of MITM attacks to reduce the risk of MITM attacks in corporations and urges organizations to encrypt their data. Additionally, this research aims to better understand the Rivest Shamir Adleman (RSA) and Elliptic Curve Cryptography (ECC) algorithms. Based on that, the medical organizations' systems may select the most appropriate ones for their needs. The proposed innovative method involves evaluating both algorithms using four performance measures on five distinct security level bits, as suggested by National Institute of Standards and Technology (NIST) [9]. The result will enhance the existing medical organizations' systems and

help the medical organizations' engineers to choose the optimal encryption algorithm.

The body of the article has the following structure: Section II includes the background and related work. Section III presents the research methodology. Section IV explains the results and discussion. Finally, the conclusions and possible guidelines for further work are presented in Section V.

II. Background and Related Work

A. Background

Modern society relies on communication networks and the Internet for almost every facet of everyday activities. Like online home banking, social media networks, and online shopping, most applications need cellular networks or the Internet. This is the major target of hackers since it involves transmitting sensitive information. Hackers prey on businesses and organizations, causing enormous financial damage [10]. The MITM attacks are the most effective method of controlling sensitive end-user information being sent. Therefore, it is one of the most serious risks to the security of wireless networks. A typical MITM attack scenario includes the victims, the two endpoints, and the perpetrators, a third party [11]

During a security breach, an attacker gets into a communication system and changes messages between the two endpoints. Third-party attackers can intercept, alter, replace, or alter information being carried across the communication channel between two endpoints when they conduct MITM attacks. Due to their lack of knowledge, victims feel that their communication channels are secure. Global System for Mobile (GSM), Universal Mobile Telecommunications System (UMTS), Bluetooth, and Wi-Fi are a few of the communication channels that may be used to execute such MITM attacks [11]. The hackers also compromise the data's security by targeting the actual data sent between the endpoints.

An adversary may tamper with the secrecy and integrity of communication [3]. Alternatively, an adversary may stop communication between the two parties and weaken the availability issue by intercepting, modifying, or destroying the messages. As shown in Fig. 1, the authors explain in how user one and user two do not have a trusted connection with the MITM.

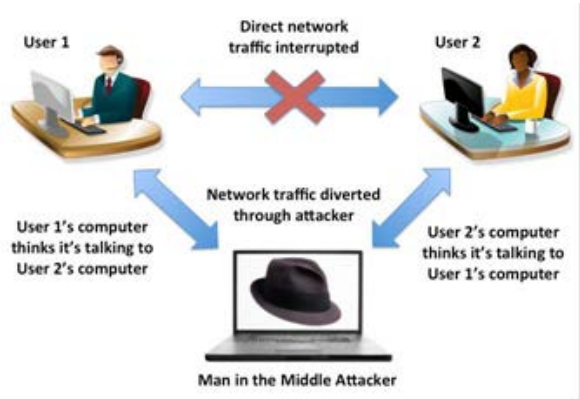


Fig. 1. Traditional MITM Attack [3].

A malicious attacker can intercept and decrypt the data passed between sensors and the Local Processing Unit (LPU) [4]. Consequently, an attacker may access confidential information and assess the recorded data to discover significant changes or clinical concerns. For example, MITM may change incorrect data and communicate normal readings to the LPU, preventing the monitoring system from sounding an alert when a patient asks for help. The author demonstrated the same malign spirit by using the Medtronic infusion pump to block it from administering insulin or overloading diabetic patients with insulin. The sensor's data is usually normal, with just a few exceptions. In [4], the authors note that the LPU analyzes the data to look for significant shifts in measurements before sounding the alert. Because the MITM cannot access personal details, the sensor merely transmits a digital signature of what it has collected. This interval between readings is preserved by the change detection mechanism in the LPU. Researchers employ Locality Sensitive Hashing (LSH) to create an irreversible information fingerprint that makes it impossible for an adversary to deduce or access confidential data.

In contrast, sending signatures rather than measurements greatly decreases the size of the data packet and, as a result, the amount of energy needed to transmit it. IMD is a trusted healthcare platform that has sensitive patient information. The authors in [11] explain that attackers can listen, change, and drop the messages when the medical unit loses authentication with the patient's IMD (Fig. 2).

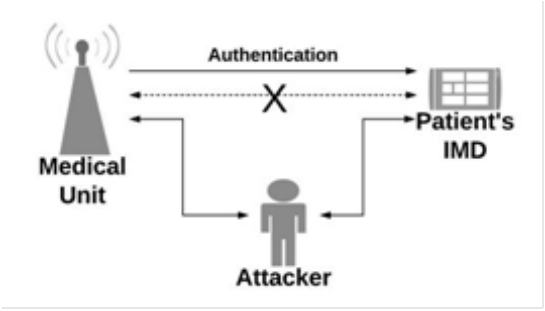


Fig. 2. Example of MITM Attacks Scheme in a Medical Unit [12].

B. Related Work

In [4], the authors reviewed aims to mitigate MITM attacks on the Internet of medical systems. Specifically, the attack happens by identifying the respective monitored individuals' healthcare emergency and replaying normal physiological data to prevent the system from raising the alarm. The authors depend on Locality Sensitive Hashing (LSH) signature as transmitted instead of physiological value. To prevent modification, replay, and black hole attacks, a Hash-Based Message Authentication Code (HMAC) is used with a key based on the Received Signal Strength Indicator (RSSI) value measured on both sensors and (Local Processing Unit) LPU. Also, propose a system that could be leveraged to prevent the devastating aftermath effects of the alarms of the remote healthcare monitoring system.

In [13], the authors proposed an efficient scheme to help design a generalized yet robust authentication protocol in medical systems. It is a countermeasure against medical

facilities' potential man-in-the-middle attacks and impersonation attacks. Specifically, the countermeasure involves mutual authentication between users, their devices, and the system's cloud server. It also involves standardizing a key agreement scheme with Elliptic Curve Cryptography (ECC). With the model in place, the authors opine that the keys are thoroughly secured, hence not copy-able; therefore, it is pivotal in enhancing security robustness.

In [14], the authors examined and proposed using a lightweight cross-layer trust computation algorithm for the MITM attacker detection, known as IC-MADS. IC-MADS are identified to have two notable contributions that the others, such as the trust-based and cryptography-based solutions, failed to possess, which relates to energy-efficient clustering and cross-layer attack detection. According to the authors, simulation results identify IC-MADS as efficient in achieving better protection against potential MIMA attacks with minimum energy consumption.

In [15], the authors have proposed a biometric-based authentication scheme that would help ensure secure access to patient's electronic health records virtually from any location. There has been a notable trend in Healthcare 4.0-based diagnostics systems globally. However, the authors often identify that patient records are continually stored in Electronic Health Records (EHR) repositories. Therefore, they use RSA encryption to protect patient data security and privacy risks. Nevertheless, results attribute the scheme as superior to the previously used state-of-the-art schemes.

In [16], the authors discussed the aspects of big data, especially in the modern-day context where it has been most impactful across industries. For example, benefits include driving health research, enhancing knowledge discovery, and improving personal health management in the healthcare domain. Primary identified big data challenges include technical challenges and privacy and security issues. The authors recommend incorporating encryption and anonymization as the best practices in enhancing big data security and privacy.

In [17], the authors identified the potential risk of the Internet of Things (IoT) era, especially with the continued advent of technology. Therefore, the security and protection of IoT depend on various factors, ranging from the producer of the device and their respective perception of device protection to the end-user and their probable awareness of the associated risks. Furthermore, in [17], the authors noted that attackers are often at an advantage concerning their inherent knowledge and technology. Therefore, despite the apparent great potential of IoT, it is faced with considerable risks that stem from insufficient protection. Therefore, advancing prevention and reactiveness is the best approach to managing the situation.

Table I shows the advantages and disadvantages of exploring related work.

TABLE I. ADVANTAGES AND DISADVANTAGES OF THE RESPECTIVE STUDIES

Related Studies	Advantages	Disadvantages
[4]	The authors successfully proposed an effective mitigation strategy to lessen the negative impact of MITM attacks on the Internet of Medical Things (IoMT). Furthermore, the approach successfully addressed critical domains, such as the privacy of the physiological data and energy consumption.	Using a classification model increases the risk of failure of the strategy.
[13]	The proposed scheme is a robust authentication protocol and can fulfill its scope within the medical infrastructure.	Mutual authentication can be disastrous, especially when parties fail to honor their pledges.
[14]	Simulation results prove that IC-MADS are integral in better protection against MITM attacks and leverage minimum overhead and energy consumption.	It is associated with a limited power rating as it is usually impossible to manufacture higher power.
[15]	The proposed biometric-based authentication method proved superior in its computational and communication costs, especially when compared to conventional schemes.	The proposed solution took more time than expected. The author suggests decreasing the encryption key.
[16]	Encryption and anonymization are unique solutions to the presenting data privacy and security challenges.	Encryption could be disadvantageous as it consumes significant resources and has issues with data compatibility.
[17]	IoT has great potential in present-day society due to increased technology and expertise. Providing certificates to identify each device will mitigate the MITM attack.	Certificates might be an insufficient solution due to the high cost. In addition, insufficient protection of the users results in increased privacy concerns.

III. RESEARCH METHODOLOGY

This chapter will debate the proposed research methodology used to achieve the research objective. Assimilation is the most popular experimentation technique in the network field. This chapter further discusses the proposed authentication scheme aimed at helping to overcome the impersonation and the MITM attacks during the user login and data storing phases, respectively. First, it introduces the authentication scheme and details how the attacks occur. It then proposes a solution for the attacks.

This research proposes a new authentication scheme for cloud computing for mobile users. The proposal is motivated by the rising levels of attacks on wireless channels. The research views that authentication and verification are critical elements that can help enhance the security channels between mobile users and cloud computing. The prosed solution in this research involves implementing two-layer security with a crucial agreement scheme. Cryptography, a well-known approach for securing communication, is proposed to be the baseline for the proposed solution.

The methodology of this research is demonstrated in Fig. 3. The figure provides an overview of the different phases of this research methodology. It begins with studying previously done literature reviews related to the research area. The focus is mainly on the literature reviews completed in recent years, whose primary study is how health care systems' phases work. After that, the focus is on the MITM attack on the health care systems. Then, the research will implement a solution that mitigates the MITM attacks. The performance of the proposed solution is then analyzed. Finally, the results are generated and discussed.

In addition, the experiment was done on a hardware server. A real server chassis model Quanta S5HF-1U was rented. To conduct the experiments, the processor is 1× AMD EPYC 7281. CPU - 16C/32T - 2.1 GHz. Storage 2 × 1 TB NVMe. 96 GB DDR4 ECC. The operating system is Ubuntu 22.04 LTS.

Furthermore, This research used python language to experiment. Python was chosen because it is a dynamically semantic, interpreted, object-oriented high-level programming language. Its high-level built-in data structures, combined with dynamic typing and binding, make it ideal for Rapid Application Development and scripting or glue language for connecting existing components.

A. Health Care System

Two phases form a basic wireless and healthcare structure. The first phase is the login phase, while the second phase is called the information storing stage. The assumption is that an impersonation attack occurs at the user-long stage. An assumption is that an attack on the man in the middle is usually in the information storing stage [13] (See Fig. 4).

The absence of authentication between a sender and receiver is one of the elements leading to an impersonation attack in a medical center. An impersonation attack involves the success of external adversaries in stealing the identity of one authorized system user or stealing a communication protocol. Therefore, an impersonation attack is bound to occur when one of the user accounts is stolen. In most instances, these attacks are usually through emails attempting to impersonate someone trustworthy. The attacks also attempt to mimic an organization to access company information and finances [18].

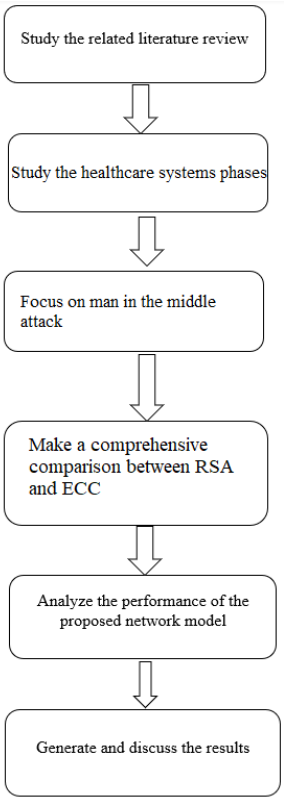


Fig. 3. Research Methodology.

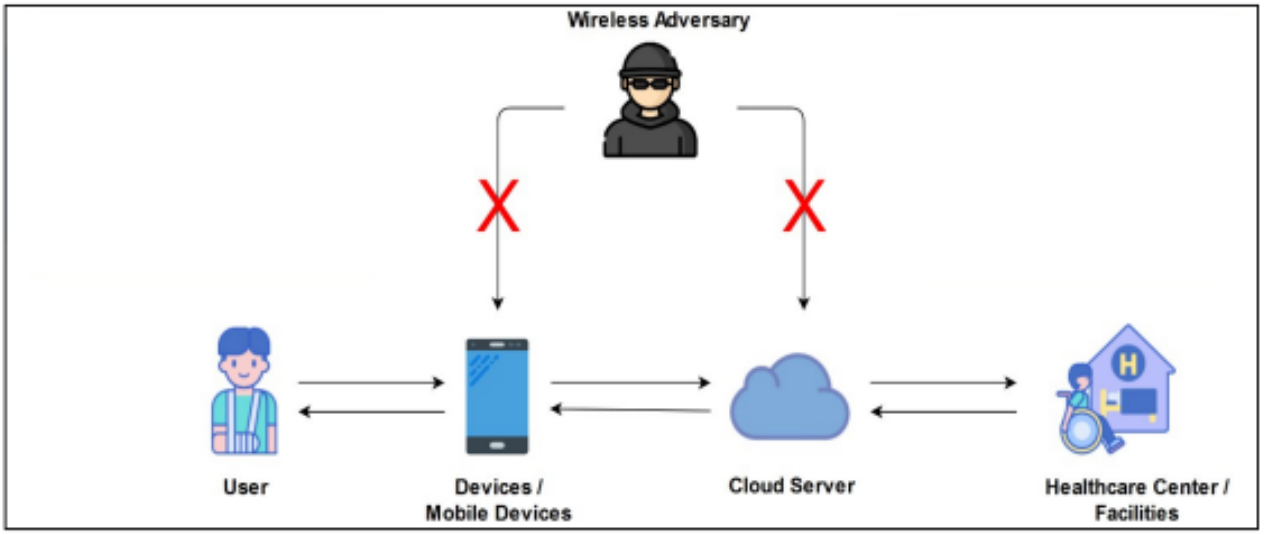


Fig. 4. Wireless Adversary Attack [13].

B. Asymmetric Keys

Asymmetric key cryptography, often known as public-key cryptography, is a type of encryption that uses asymmetric keys. In cryptography, keys are divided into two types: the first is a public key used for encryption, and the second is a private key used for decryption, as shown in Fig. 5 and Fig. 6. A certain user or device can only access the private key. Nevertheless, on the other hand, the public key is disseminated to all users and devices participating [19].

The speed and security strength are the most significant shortcomings of asymmetric ciphers; they are significantly slower than symmetric algorithms and more prone to intruder attacks, making the key exchange process more difficult. The advantage of using an asymmetric key technique is that it eliminates the need to distribute the encryption key between parties. Private keys are kept secret, and only public keys are made available to the public [19].

In addition, Digital signatures are possible with public key encryption, allowing the communication recipient to verify that the message came from the sender who specified the digital signature. With digital signatures in public-key encryption, the receiver can determine whether or not the message has been altered during transit. No changes can be made to a digitally signed communication without invalidating the signature [19].

If part A wants to communicate with part B confidentially, it should encrypt a message using B's publicly available key. Because only B has access to the associated private key, such communication can only be deciphered by B.

If part A wants to send an authenticated message to party B, as shown in Fig. 7, part A should encrypt the message using A's private key. Because this message can only be deciphered using A's public key, which may be used to verify the message's authenticity, A is indeed the message's source [20].

At the same time, public-key cryptography may support message authentication and confidentiality. For example, Fig. 8 shows how public-key cryptography ensures authentication.

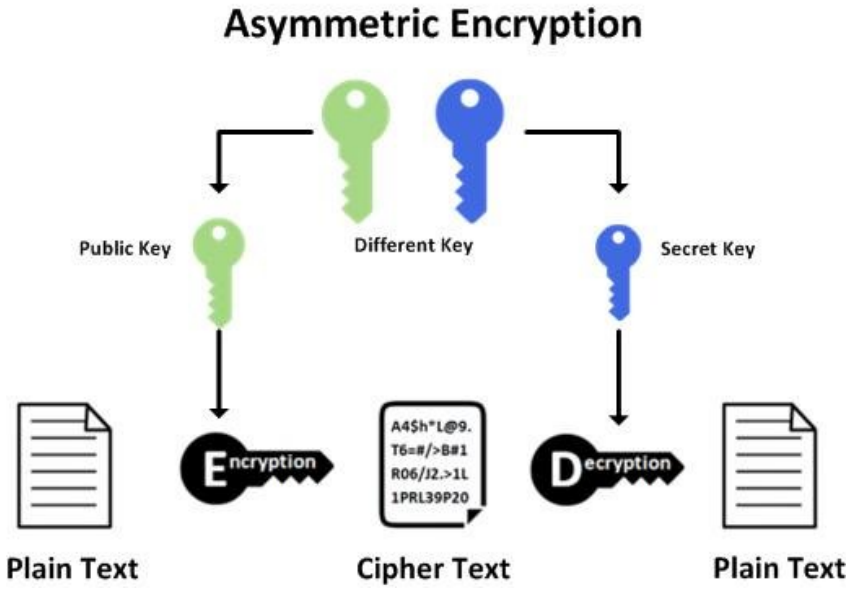


Fig. 5. Asymmetric Encryption [19].

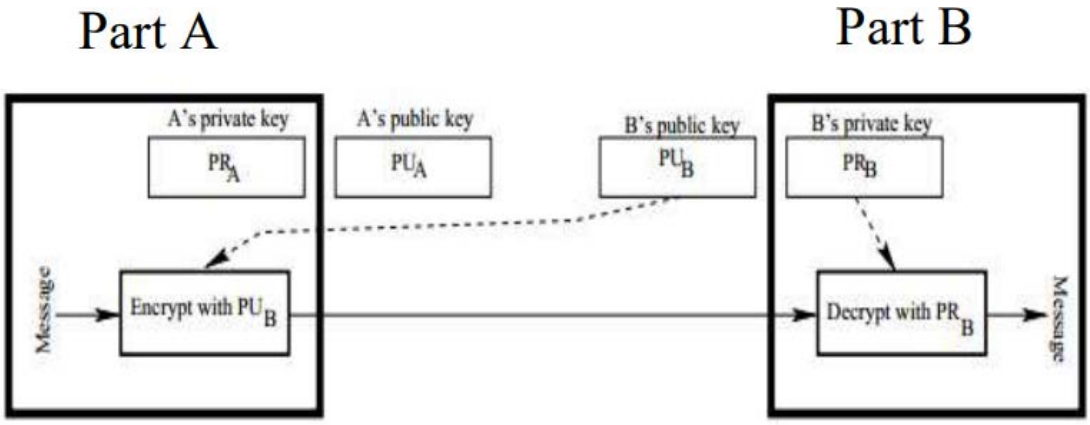


Fig. 6. Confidential Communication in Public-Key Cryptography [20].

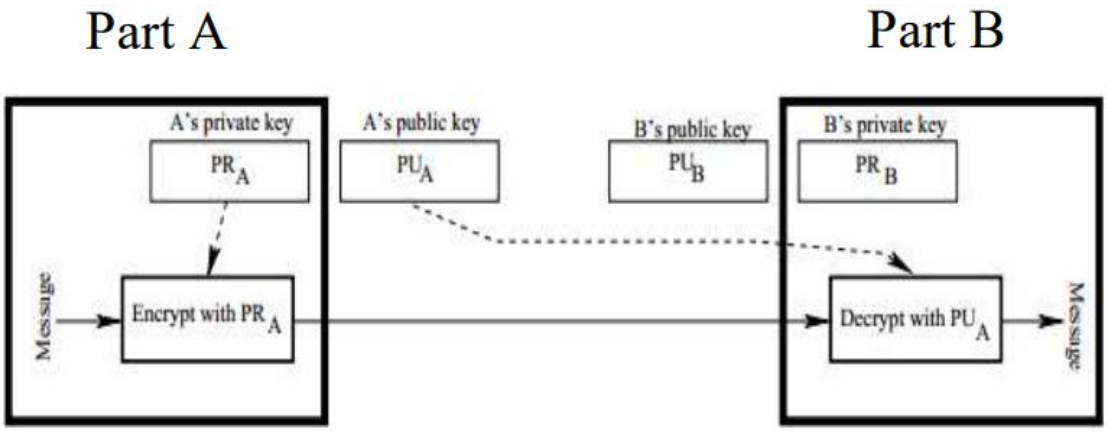


Fig. 7. Authenticated Communication in Public-Key Cryptography [20].

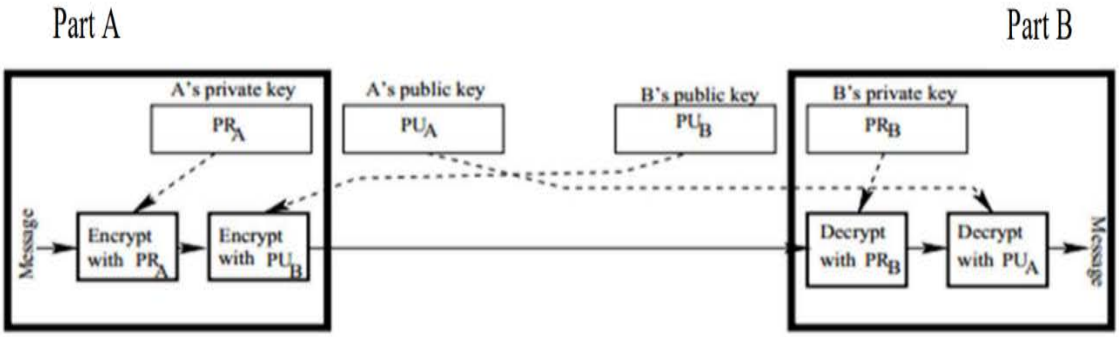


Fig. 8. Confidentiality and Authentication [20].

Fig. 8 illustrates how public-key cryptography can be used for confidentiality and authentication, including digital signatures. RSA and ECC can provide security services such as Confidentiality, Integrity, Authentication, and Authorization. Authors in [20] defined them as below as;

1) Confidentiality: Any illegal connection to the data via this security service is refused.

2) Integrity: Ensuring that messages sent to a destination have not been tampered.

3) Authentication: Any anonymous/malicious node wishes to interact with network nodes. It needs the authorized node's public key pair.

4) Authorization: This service assigns each node a unique key pair (private and public) for decryption and encryption.

C. RSA

The RSA algorithm, named after its creators, is the first method used for data encryption and digital signatures simultaneously. It is the most widely used today. The RSA algorithm's security depends on how difficult it is to decompose large integers. The public and private keys are created using two huge prime integers employed to generate the public and private keys. A rough estimate of how difficult it is to deduce the plaintext from the signal key and the ciphertext is the decomposition of the product of two large prime numbers [21].

In the Internet Key Exchange (IKE) architecture, the RSA algorithm has been proposed as a potential authentication technique. The Diffie-Hellman key exchange method is critical to the framework's security architecture. Participants interact using the Diffie-Hellman algorithm and create shared keys at the start of a key agreement session. These shared keys will be utilized for the key agreement protocol of the next steps [21].

Encrypting with private or public keys provides RSA users with many services. If the public key is used for encryption, the data must be decrypted with the private key. This is ideal for delivering sensitive data over a network or over the Internet, where the data recipient sends the data sender their public key. The data sender then encrypts the sensitive information with the recipient's public key and sends it to them. The private key owner can only decrypt the sensitive data because the public key encrypts it. Thus, even if the data is intercepted in transit, only the intended recipient can decode it. Fig. 9 explains how RSA encryption works [21].

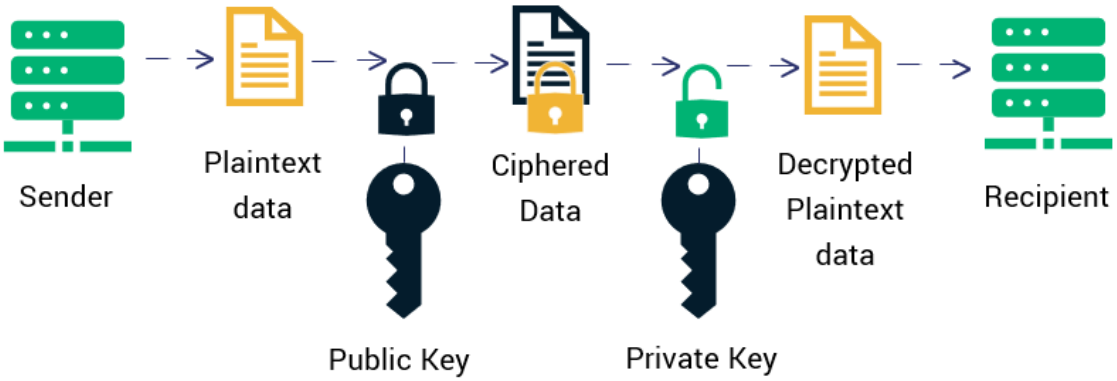


Fig. 9. How RSA Works [21].

Encrypting a message with a private key is the other asymmetric encryption method with RSA. In this case, the data sender encrypts the data with their private key and sends the encrypted data together with their public key to the data recipient. The recipient can then decrypt the data using the sender's public key, proving that the sender is whom they say they are. The data could be stolen and read in transit using this method. However, the primary purpose of encryption is to prove the sender's identity. The public key would be unable to decrypt the new message if the data was stolen and modified in route, and the recipient would be aware that the data had been altered in transit [21].

The technical aspects of RSA are based on the premise that it is simple to construct a number by multiplying two sufficiently large numbers together. Still, it is incredibly difficult to factorize that number back into the original prime numbers. For example, two numbers are used to construct the public and private keys, one of which is a product of two huge prime numbers. To calculate their value, they both use the same two prime numbers [21].

RSA is widely regarded as the first real-world asymmetric-key cryptosystem. For public-key cryptography, it becomes the de-facto standard. The integer factorization problem guarantees its safety. However, the decryption technique used by RSA is less efficient than the encrypting process. Many scholars have advocated using the Chinese Remainder Theorem (CRT) to improve the efficiency of RSA decryption. Authors in [21] suggested a CRT model improves RSA decryption time. They also advocated using a small matrix order to obtain big modulus and cryptographic keys. Larger key sizes are required for better and stronger data security, which involves higher overhead on computer systems. Small gadgets are becoming increasingly vital in today's digital world, with less memory but need security to meet market demand. RSA becomes a secondary consideration in this case.

RSA Algorithm

Key Generation
Step 1. Select p, q where p and q both are primes, $p \neq q$
Step 2. Calculate $n = pq$
Step 3. Calculate $\Phi(n) = (p-1)(q-1)$
Step 4. Select integer e $gcd(\Phi(n), e) = 1$; where $1 < e < \Phi(n)$
Step 5. Calculate d ; $d \equiv e^{-1} (mod\ \Phi(n))$
Step 6. Public key = {e, n}
Step 7. Private key = {d, n}

Encryption
Step 1. Plaintext: M<n
Step 2. Ciphertext: $C=M^e \bmod n$

Decryption
Step 1. Ciphertext: C
Step 2. Plaintext: $M=C^d \bmod n$

Each party must generate its keys to communicate safely with one another. First, the value of e in the RSA algorithm for encryption should be chosen so that gcd((n), e) equals 1. Once e has been chosen, the appropriate 'd' for decryption should be constructed by determining the inverse of 'e' mod n. During the encryption process, a sender must encrypt the message, i.e., in decimal digits, using the receiver's public key, i.e., e and n. The recipient must decrypt the ciphertext using his private key, represented by the letters d and n.

D. ECC

The ECC algorithm is public-key cryptography (PKC) with public and private keys for authentication. ECC is known as a sort of PKC built upon the algebraic structure of the elliptic curve over finite fields. The difficulty of the elliptic curve discrete logarithm problem (ECDLP) plays a major role in the security of ECC, and this problem can be resolved exponentially. Meanwhile, it has to be added that the performance of this algorithm is mainly intertwined with the efficiency of its scalar multiplication algorithm. Hamming weight of the private key is a determinant factor in algorithm efficacy regarding the scalar arithmetic level of the computation. Hamming weight measures the number of non-zero digits in a scalar representation. As the extent of Hamming's weight lowers, the speed of scalar multiplication performance rises. Accordingly, the scalar recoding method can be used to lessen the Hamming weight of the private key's scalar representation [22].

Because of its lower-key size and capacity to preserve security, ECC has gradually gained popularity over the last several years. Due to the increasing size of keys and the increasing desire for devices to remain secure, this trend will likely continue as mobile resources become more precious and the demand for devices to remain secure increases. To fully

comprehend elliptic curve cryptography, it is necessary to understand it in context. It also makes sense to use ECC to maintain high performance and security levels [22].

The ECC is becoming increasingly popular as businesses attempt to improve the online security of client data and the mobile optimization of their sites simultaneously. As the number of sites that use elliptic curve cryptography to secure data grows, the demand for brief guides to elliptic curve cryptography also grows. For the current ECC, an elliptic curve is a plane curve over a finite field composed of the points meeting the following equation: $y^2=x^3 + ax + b$. as shown in Fig. 10. It is possible to mirror any point on this elliptic curve cryptography example over the x-axis and yet have the curve retain its shape in this example. Any non-vertical line will intersect the curve three times or less if it is not vertical [22].

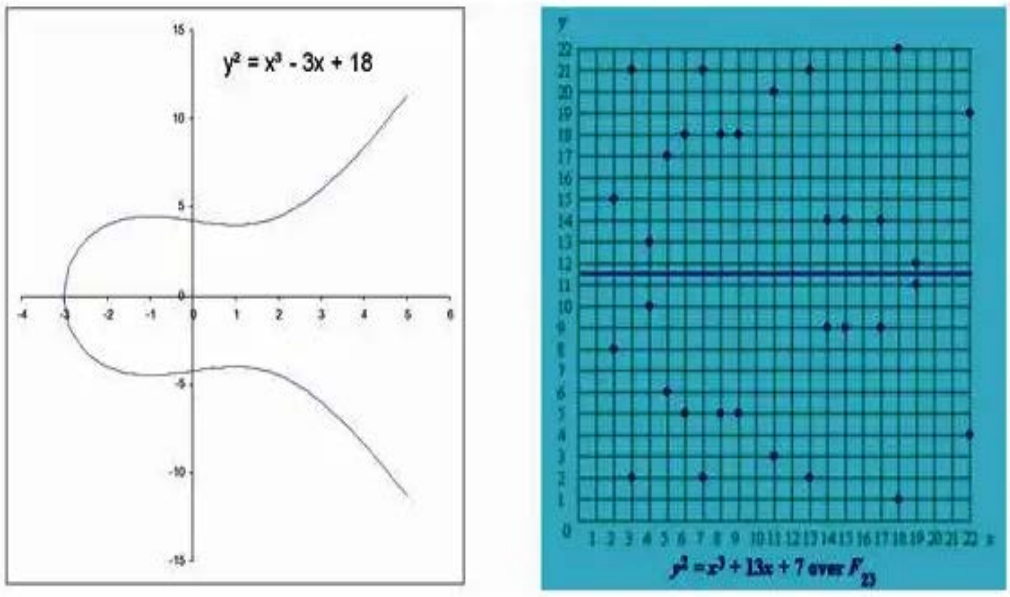


Fig. 10. 3rd-Degree Elliptic Curves [22]

Authors in [22] coined ECC as another potential asymmetric key cryptosystem in the late 1980s. This type of technology is best suited for devices with limited memory, such as Palmtops, Smartphones, and Smartcards. An ECC requires fewer or smaller parameters than RSA for encryption and decryption but with equal degrees of security.

ECC Algorithm

Global Public Elements
Step 1. $E_q(a, b)$ elliptic curve with parameters a, b, and q, where q is a prime or integer of from 2^m.
Step 2. G point on the elliptic curve whose order is large value n, where n is the mod.

User Alice Key Generation
Step 1. Select private key n_A; where $n_A < n$
Step 2. Calculate public key P_A
Step 3. $P_A = n_A G$

User Bob Key Generation
Step 1. Select private key n_B; where $n_B < n$
Step 2. Calculate public key P_B
Step 3. $P_B = n_B G$

Calculation of Secret Key by User Alice
Step 1. $K = n_A P_B$
Calculation of Secret Key by User Bob
Step I. $K = n_B P_A$

Encryption by Alice using Bob's Public Key
Step 1. Alice chooses the message P_m and a random positive integer k.
Step 2. Ciphertext: $C_m = \{ kG, P_m + kP_B \}$

Decryption by Bob using his own Private Key
Step 1. Ciphertext: C_m
Step 2. Plaintext: $P_m = P_m + kP_B - n_B (kG) = P_m + k(n_B G) - n_B (kG)$
P_m is a (x,y) point encoded with the plaintext message m in this case. Encryption and decoding take place at the P_m.

E. Performance Metrics

This section of the paper will determine the performance metrics that have been used to base our comparison of the RSA and ECC algorithms on. Performance metrics can take on various forms, but the focus is on just four types for this research.

1) Memory utilization

Memory is an integral component of the entire computer system and essentially consists of a system of devices that helps in data storage on electronic digital computers. Computer memory can either be temporary or permanent, although this depends largely on the frequency of data retrieval [23]

Memory utilization is calculated by storing the resident set size before the encryption or decryption functions. After running the encryption or decryption functions, the system time is stored in another variable. Now the difference between these two variables is memory utilization.

2) Signature generation time

When sending data, for instance, through a document, it is paramount to identify the authenticity of the senders, for optimal security and safety, for instance, against the distinct forms of cyber theft [24].

Signature generation time is calculated by storing the system time value in a variable, then running the generation function. After that, store the system time in another variable. The difference between these two variables is the signature generation time.

3) Signature verification time

The use of signature verification by algorithms means an effort to unearth the identity of the parties involved in sending and receiving messages and is integral in facilitating timely identification and aversion of potential threats that could negatively affect data security and integrity [24].

Signature verification time is calculated by storing the system time value in a variable, then running the verification function. After that, store the system time in another variable. The difference between these two variables is the signature verification time.

4) Encryption and decryption time

Encryption time is required to convert plaintext to ciphertext, while decryption time is required to convert ciphertext to plaintext [25].

Encryption and decryption time is calculated by storing the system time value in a variable before the encryption. Then, running the encryption or decryption function. After that, store the system time in another variable. Now, the difference between these two variables, encryption or decryption time

F. NIST Recommendation

The comparable key-size classes addressed in this section are based on estimations generated using currently available methodologies as of the publishing of this Recommendation. Future advancements in factoring algorithms, general discrete-logarithm assaults, elliptic-curve discrete logarithm attacks, and quantum computing may impact these equivalencies. In addition, new or improved attacks or technologies may emerge, rendering some of the current methods utterly insecure. For example, if quantum attacks become realistic, asymmetric approaches may no longer be secure. Periodic reviews will be conducted to see if the stated equivalencies need to be altered. For example, key sizes need to be increased or if the algorithms are no longer secure. Other than brute-force cryptographic attacks, strong cryptographic algorithms may be able to mitigate security vulnerabilities. For example, the algorithms may be built, so those small quantities of information about the key are unintentionally leaked. In this situation, the larger key may lower the chances of a compromised key due to the disclosed information [26]. Table II shows equivalent maximum-security strengths for the accepted algorithms and key lengths.

TABLE II. NIST RECOMMENDED SECURITY BIT LEVEL (BARKER, 2020)

Security Bit Level	RSA	ECC
80	1024	160
112	2048	224
128	3072	256
192	7680	384
256	15360	512

Security bit level is a cryptographic primitive's security level measures its strength, such as a cipher or hash function. The security level is commonly stated in bits, with n-bit security implying that breaking it would take 2n operations [26].

IV. RESEARCH RESULTS AND DISCUSSIONS

This chapter contains the analysis comparison parts of RSA and ECC algorithms based on the security bit-level suggested by NIST. Memory utilization is in bytes, signature generation time, signature verification, encryption, and decryption time are in milliseconds.

A. Memory Utilization

Based on Fig. 11 and Table. III, ECC shows better than RSA in memory utilization at all security bit levels; ECC needs less memory usage than RSA [20]. A massive spike after 192-bit level in RSA was observed. That makes RSA worst in memory handling, especially in the large keys.

TABLE III. MEMORY UTILIZATION COMPARISON BETWEEN RSA AND ECC

Security Bit Level	Memory Utilization in bytes	
	RSA	ECC
80	160	109
112	239	119
128	315	127
192	620	144
256	1777	220

B. Signature Generation Time

Fig. 12 and Table. IV show that ECC and RSA are close to each other in the signature generation time. RSA is better at 80, 112, and 128 security bit levels. In the 192-security bit-level, a small RSA latency compared to ECC was observed. Also, great latency in RSA at the 256-security level was noticed. RSA needs 3 ECC times to generate the signature [27].

TABLE IV. SIGNATURE GENERATION TIME COMPARISON BETWEEN RSA AND ECC

Security Bit Level	Signature Generation Time in Milliseconds	
	RSA	ECC
80	0.0102	0.1530
112	0.1533	0.3411
128	0.2119	0.5912
192	1.5322	1.1897
256	9.2152	3.087

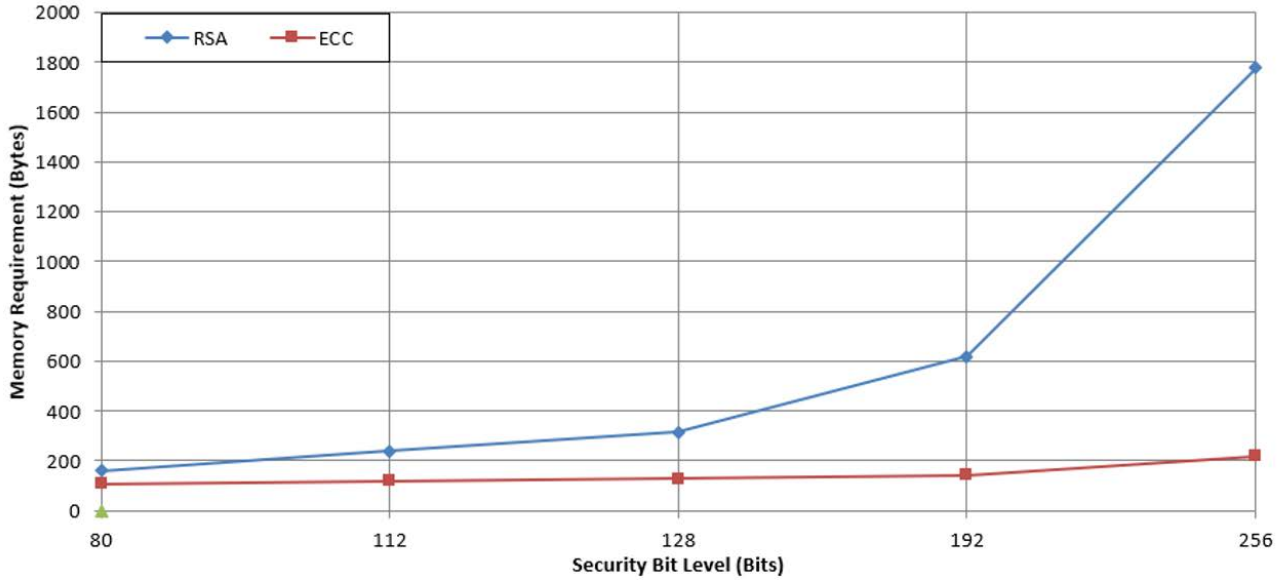


Fig. 11. Memory Utilization Comparison Graph between RSA and ECC

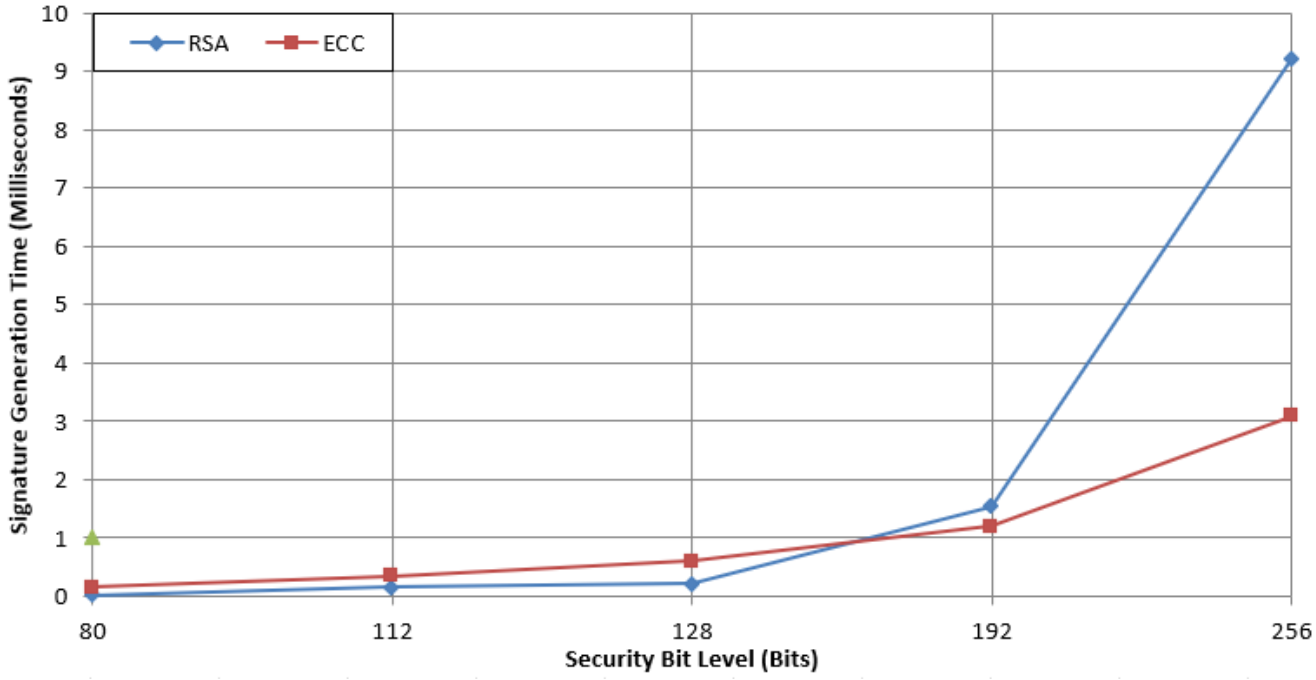


Fig. 12. Signature Generation Time Comparison Graph between RSA and ECC

C. Signature Verification Time

As shown in Fig. 13 and Table. V, RSA trumps the ECC in all security bit levels regarding signature verification. In RSA, the time required to verify a signed message is trivial for the key length employed. However, ECC is significantly slower to perform in each key range and exhibits an almost linear rise in performance with increasing the security bit level [20]. The reason because in RSA, the cost of verification can be controlled to be minimal.

TABLE V. SIGNATURE VERIFICATION COMPARISON BETWEEN RSA AND ECC

Security Bit Level	Signature Verification Time in a millisecond	
	RSA	ECC
80	0.0110	0.2310
112	0.0116	0.5231
128	0.0124	0.8622
192	0.0130	1.8100
256	0.0310	4.5410

D. Encryption and Decryption Time

Fig. 14 and Table VI show that RSA is very fast compared to ECC. in all security bit levels. Even with 256 bits, RSA needs around 1.03 seconds for encryption [28].

TABLE VI. ENCRYPTION TIME COMPARISON BETWEEN RSA AND ECC

Security Bit Level	Encryption Time in seconds	
	RSA	ECC
80	0.0306	0.4886
112	0.0310	2.2030
128	0.0360	3.8763
192	0.0489	5.2113
256	1.0310	8.5441

Fig. 15 and Table VII show a noticeable massive RSA change when the security bit level increases. Only on 80 bit of security level than RSA is better in decryption time. When the security bit level increments, a high time increment in RSA was observed. After 80 security level bit, ECC becomes better than RSA.

TABLE VII. DECRYPTION TIME COMPARISON BETWEEN RSA AND ECC

Security Bit Level	Decryption Time in seconds	
	RSA	ECC
80	0.7634	1.3376
112	2.7165	1.6012
128	7.1022	1.7770
192	14.002	2.0031
256	22.120	4.1194

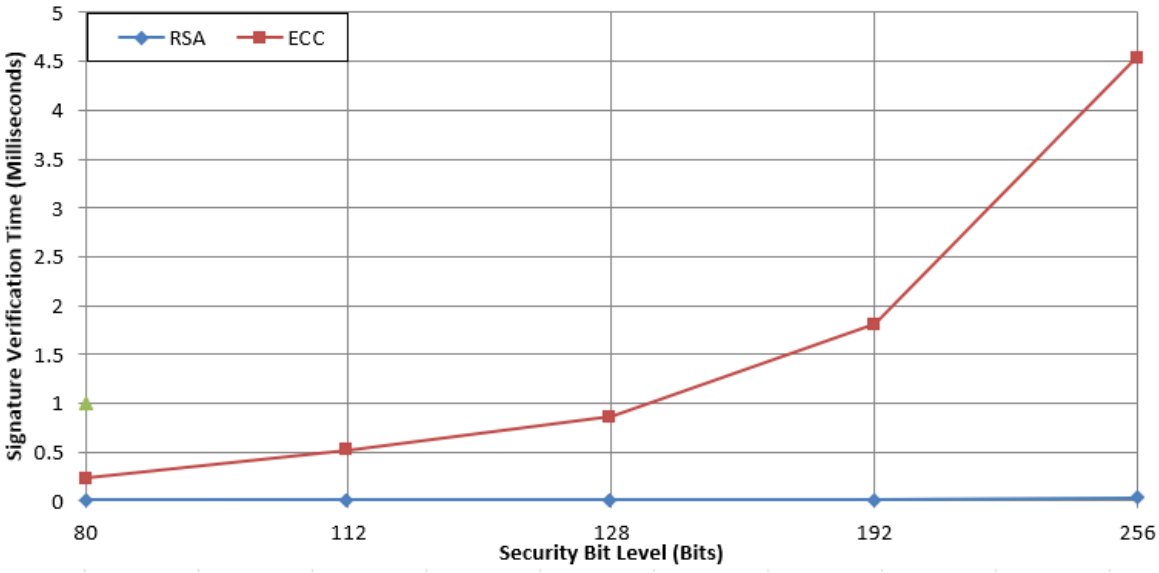


Fig. 13. Signature Verification Comparison Graph between RSA and ECC

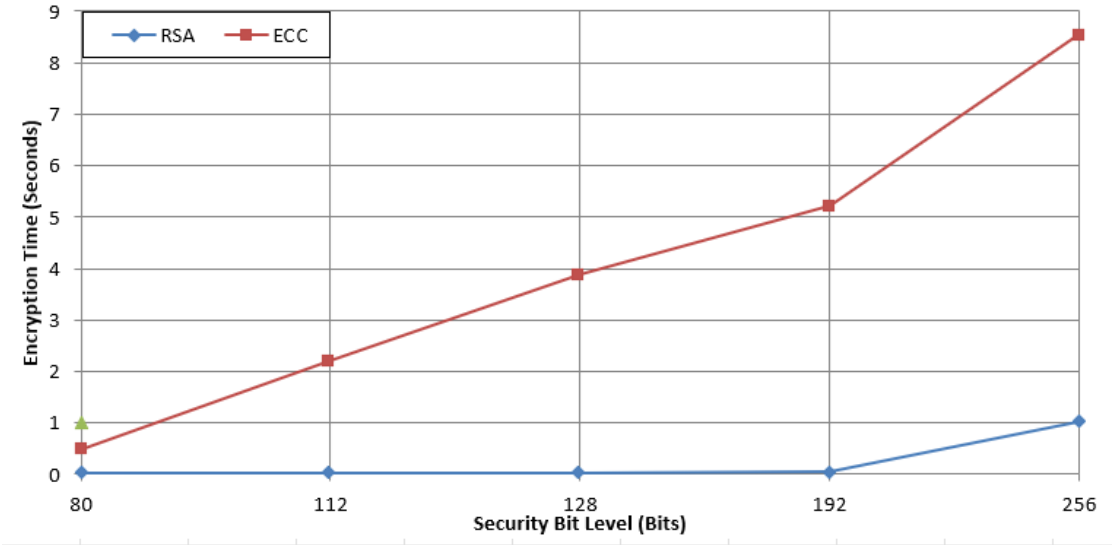


Fig. 14. Encryption Time Comparison Graph between RSA and ECC

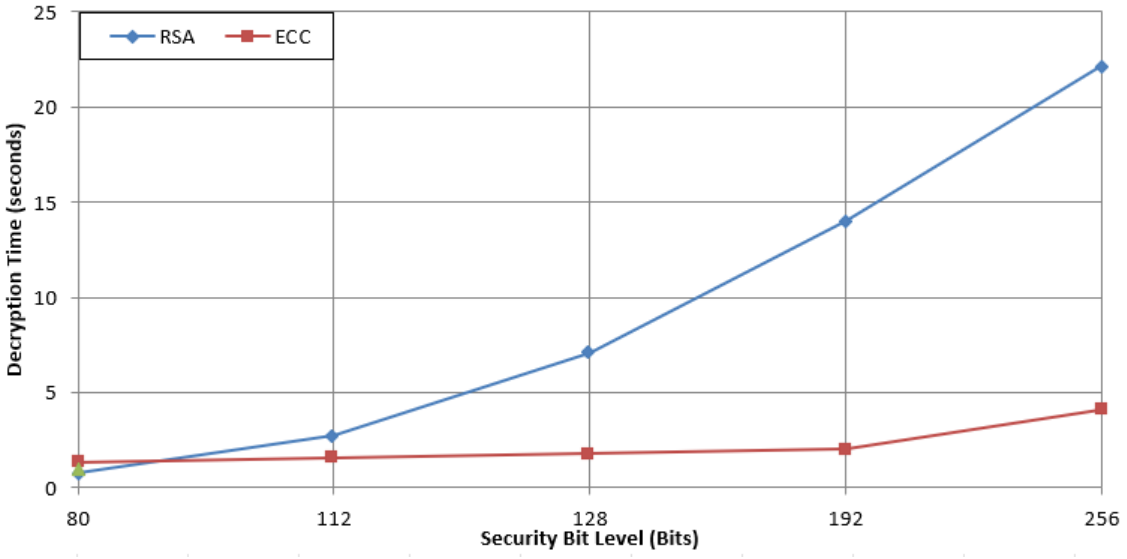


Fig. 15. Decryption Time Comparison Graph between RSA and ECC

V. Conclusion and Future Work

The majority of organizations and individuals rely on asymmetric encryption algorithms despite their complexity since they are safe and tough to break. Both RSA and ECC are instances of powerful asymmetric algorithms. This research analyzed the similarities and differences in both algorithms. Memory utilization, signature generation time, signature verification time, encryption time, and decryption time were used as performance metrics. The findings of this research show that ECC is more successful in memory use across all of the security bit-levels recommended by NIST. In addition, regarding the time required to generate signatures, RSA is more efficient than ECC when the security level is 80 or 112. On the other hand, when there is a rise in the security bit level, ECC becomes faster than RSA. When it comes to signature verification time, RSA is outstandingly fast, but ECC takes more than ten times as long as RSA, at the very least were used as a performance metric. RSA maintains its encryption time speed even when the security bit-level increases in encryption time. However, regarding decryption time, RSA becomes faster only when the security bill level increases by more than 80 security level bits. Although this experiment is done in a dedicated physical server, that service provider has limitations. One of them is in displaying the server's power consumption. That limits us from calculating the power consumption comparison between RSA and ECC in the five-bit security levels. Since electricity has become a primary factor in operational costs, adding a power-consuming as a new comparison parameter between RSA and ECC is the potential for future work.

References

[1] Petrov, I. Beliak, A. Kryuchyn, and A. Shikhovets, "Analysis of methods for creating media for long-term data storage," in 2020 IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT), 2020, pp. 238-241.

[2] Tabrizchi and M. Kuchaki Rafsanjani, "A survey on security challenges in cloud computing: issues, threats, and solutions," The journal of supercomputing, vol. 76, pp. 9493-9532, 2020.

[3] S. Akter, S. Chellappan, T. Chakraborty, T. A. Khan, A. Rahman, and A. A. Al Islam, "Man-in-the-middle attack on contactless payment over NFC communications: design, implementation, experiments and detection," IEEE Transactions on Dependable and Secure Computing, vol. 18, pp. 3012-3023, 2020.

[4] O. Salem, K. Alsubhi, A. Shaafi, M. Gheryani, A. Mehaoua, and R. Boutaba, "Man-in-the-Middle attack mitigation in internet of medical things," IEEE Transactions on Industrial Informatics, vol. 18, pp. 2053-2062, 2021.

[5] M. Z. Gunduz and R. Das, "Cyber-security on smart grid: Threats and potential solutions," Computer networks, vol. 169, p. 107094, 2020.

[6] T. Hidayat and R. Mahardiko, "A Systematic literature review method on aes algorithm for data sharing encryption on cloud computing," International Journal of Artificial Intelligence Research, vol. 4, pp. 49-57, 2020.

[7] P. Brandão, "The importance of authentication and encryption in cloud computing framework security," International Journal on Data Science and Technology, vol. 4, pp. 1-5, 2018.

[8] S. Kumar, P. Tiwari, and M. Zymbler, "Internet of Things is a revolutionary approach for future technology enhancement: a review," Journal of Big data, vol. 6, pp. 1-21, 2019.

[9] S. Greenberg, L. P. Mason, S. O. Sadjadi, and D. A. Reynolds, "Two decades of speaker recognition evaluation at the national institute of standards and technology," Computer Speech & Language, vol. 60, p. 101032, 2020.

[10] M. Knežević, S. Tomović, and M. J. Mihaljević, "Man-in-the-middle attack against certain authentication protocols revisited: Insights into the approach and performances re-evaluation," Electronics, vol. 9, p. 1296, 2020.

[11] B. Bhushan, G. Sahoo, and A. K. Rai, "Man-in-the-middle attack in wireless and computer networking—A review," in 2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA)(Fall), 2017, pp. 1-6.

[12] T. Belkhouja, A. Mohamed, A. K. Al-Ali, X. Du, and M. Guizani, "Light-weight solution to defend implantable medical devices against man-in-the-middle attack," in 2018 IEEE Global Communications Conference (GLOBECOM), 2018, pp. 1-5.

[13] S. O. Maikol, A. S. Khan, Y. Javed, A. L. A. Bunsu, C. Petrus, H. George, et al., "A novel authentication and key agreement scheme for countering MITM and impersonation attack in medical facilities," International Journal of Integrated Engineering, vol. 13, pp. 127-135, 2021.

[14] Kore and S. Patil, "IC-MADS: IoT enabled cross layer man-in-middle attack detection system for smart healthcare application," Wireless Personal Communications, vol. 113, pp. 727-746, 2020.

[15] J. J. Hathaliya, S. Tanwar, S. Tyagi, and N. Kumar, "Securing electronics healthcare records in healthcare 4.0: A biometric-based approach," Computers & Electrical Engineering, vol. 76, pp. 398-410, 2019.

[16] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, "Big healthcare data: preserving security and privacy," Journal of big data, vol. 5, pp. 1-18, 2018.

[17] Z. Cekerevac, Z. Dvorak, L. Prigoda, and P. Cekerevac, "Internet of things and the man-in-the-middle attacks–security and economic risks," MEST Journal, vol. 5, pp. 15-25, 2017.

[18] S. Sengupta, A. Chowdhary, A. Sabur, A. Alshamrani, D. Huang, and S. Kambhampati, "A survey of moving target defenses for network security," IEEE Communications Surveys & Tutorials, vol. 22, pp. 1909-1941, 2020.

[19] M. B. Yassein, S. Aljawarneh, E. Qawasmeh, W. Mardini, and Y. Khamayseh, "Comprehensive study of symmetric key and asymmetric key encryption algorithms," in 2017 international conference on engineering and technology (ICET), 2017, pp. 1-7.

[20] Z. Vahdati, S. Yasin, A. Ghasempour, and M. Salehi, "Comparison of ECC and RSA algorithms in IoT devices," Journal of Theoretical and Applied Information Technology, vol. 97, 2019.

[21] G. Amalarethinam and H. Leena, "Enhanced RSA algorithm with varying key sizes for data security in cloud," in 2017 World Congress on Computing and Communication Technologies (WCCCT), 2017, pp. 172-175.

[22] M. A. Khan, M. T. Quasim, N. S. Alghamdi, and M. Y. Khan, "A secure framework for authentication and encryption using improved ECC for IoT-based medical sensor data," IEEE Access, vol. 8, pp. 52018-52027, 2020.

[23] N. Markham and G. Pereira, "Experimenting with algorithms and memory-making: Lived experience and future-oriented ethics in critical data science," Frontiers in big Data, vol. 2, p. 35, 2019.

[24] S. Abd Elminaam, H. M. Abdual-Kader, and M. M. Hadhoud, "Evaluating The Performance of Symmetric Encryption Algorithms," International Journal of Network Security,, vol. 10, pp. 213-319, 2010.

[25] O. P. Verma, R. Agarwal, D. Dafouti, and S. Tyagi, "Notice of Violation of IEEE Publication Principles: Peformance analysis of data encryption algorithms," in 2011 3rd International Conference on Electronics Computer Technology, 2011, pp. 399-403.

[26] Barker, W. Barker, W. Burr, W. Polk, and M. Smid, "Recommendation for key management part 1: General (revision 5)," NIST special publication, pp. 800-57, 2020.

[27] Pharkkavi and D. Maruthanayagam, "TIME COMPLEXITY ANALYSIS OF RSA AND ECC BASED SECURITY ALGORITHMS IN CLOUD DATA," International Journal of Advanced Research in Computer Science, vol. 9, 2018.

[28] Mallouli, A. Hellal, N. S. Saeed, and F. A. Alzahrani, "A survey on cryptography: comparative study between RSA vs ECC algorithms, and RSA vs El-Gamal algorithms," in 2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), 2019, pp. 173-176.

A Smart Decision Making System for the Optimization of Manufacturing Systems Maintenance using Digital Twins and Ontologies

ABADI Mohammed[1], ABADI Chaimae[2], BEN-AZZA Hussain[4]
Laboratory of Mechanics, Mechatronics and Command
Modeling, Information Processing and Control of Systems (MTICS) Team
Ecole Nationale Supérieure d'Arts et Métiers (ENSAM-Meknès)
Moulay Ismail University, B.P. 4042, 50000
Meknes, Morocco

ABADI Asmae[3]
Euromed Research Center
Euromed University of Fez
Fez,
Morocco

***Abstract*—Now-a-days manufacturing processes are becoming more and more complex which constantly complicate the management of their life cycle. Although, in order to survive and maintain a good position in the competitive industrial context, industrials have understood that they must optimize the whole life cycle of their manufacturing processes. The maintenance constitutes one of the key processes indispensable to ensure the proper functioning and to optimize the lifetime of machines and production lines, and thus to optimize quality and production costs. Therefore, its automation and optimization represent until now a center of interest for researches and manufacturers, especially those related to the integration of artificial intelligence tools in the industry. In this context, several new concepts and technologies have emerged, particularly in the context of industry 4.0. One of these new concepts is digital twins, which has become a promising direction to optimize manufacturing processes lifecycle. However, the implementation of this technology faces several complex problems related to the interoperability between physical entities and their virtual counterparts, as well as to the logical reasoning between the different elements constituting the digital twin. It is in this context that an approach based on digital twins and ontologies is proposed. The originality of this paper lies in two important points: the first is the exploitation of the expressiveness and reasoning capabilities of ontologies to solve cyber-physical interoperability problems at the digital twin level, while the second is the automation of the whole maintenance process and its decision making key points using the inference potentialities of ontologies. The applicability and effectiveness of the proposed approach is validated through an industrial case of study.**

Keywords*—Maintenance systems; maintenance policy; digital twin; reasoning; ontologies; automation; cyber-physical interoperability; decision making; artificial intelligence

I. Introduction

Mastering the maintenance process of industrial systems has become a necessity for companies, in order to pursue the continuous growth of competitive markets. The achievement of this goal will allow manufacturers to optimize the productivity and quality of their production systems, and therefore gain in terms of costs, quality and delays. To this end, industrials and researchers have started to automate this key process, but until now a complete automation is not yet achieved [1]. That justifies the first objective of this paper, which is the development of a new global approach for the maintenance process automation (MMSDTO), from the data collection phase to the establishment of maintenance plans. This operation will be based on a very important artificial intelligence tool which is the Digital Twin. In fact, it will allow collecting the necessary information related to the fields, to follow the production process in real time and to locate and predict failures in the machines. However, the use of this technology requires the resolution of the cyber-physical interoperability problem, which has become the focus of many research works. Therefore, this point constitutes the second objective to be achieved through this paper. Consequently, a new concept will be integrated in the approach, namely ontologies. In fact, their expressiveness and reasoning capabilities will be exploited to preserve the semantics of the large quantities of data exchanged between the physical and virtual spaces of the Digital Twin, to overcome the problem of interoperability and also to make it able to do the logical reasoning and generate the desired results.

Thus, the first part of this paper is a literature review containing the different concepts and key points related to the work realized, notably the Digital Twin and ontologies, as well as the limitations of previous research works and the problems that need to be overcome. Then, the operating system of the newly developed MMSDTO approach will be presented and explained. The different steps of the proposed methodology will be detailed in the following sections. Finally, the approach will be applied on an industrial case study to validate its reliability and efficiency.

II. Related Work

The digital twin is a concept that has recently been the focus of several research studies [2], especially in relation to Industry 4.0 [3]. Its appearance dates back to 2003, when Michael Grieves and John Vickers participated in a conference on product life cycle management [4]. At this event, they presented the Digital Twin as a mirror space model that is used to represent physical entities in a virtual space [5]. In fact, the

Digital Twin numerically reproduces the operation and behavior [6] in real time [7] of physical elements. Therefore, several decisions will be taken in order to optimize the production system and its productivity [6]. This justifies the proposition of the digital twin's standard structure given by Michael Grieves and John Vickers. In fact, they proposed to generalize the structure (physical entity, virtual counterpart, connection between physical and virtual spaces) on digital twins [8]. Afterwards, this structure was extended, and thus, two other elements (services and digital twin data model) were added [9] to make the structure more complete and efficient. Then, this structure was projected on the production workshops by [10]. They proposed a conceptual model with four elements [11], namely:

- Physical Shop-floor: it consists of production lines, production materials and tools, products and employees [10].
- Virtual Shop-floor: it is a virtual reproduction of the functioning of the physical shop-floor and the behavior of its elements [11].
- Shop-floor Service System: it is a set of computer tools (information systems, computer aided tools, etc.) that form the services necessary to execute the commands preventing from the physical and virtual spaces [10].
- Shop-floor Digital Twin Data: data collected from physical and virtual spaces, as well as information generated from the methods of modelization, optimization and prediction of the service space are integrated [10].

The digital twins have been used in different domains such as design [12], logistics management [13], production management [14], maintenance [15], etc. Several works have focused on maintenance. The author [16] realized a literature review on the different papers produced on digital twins for maintenance. Some of these works such as [17], [18] and [19], propose the use of digital twins to do specific maintenance tasks. There are also more general approaches to predict the asset state, in order to predict accordingly the corresponding maintenance plan [20], [21], [22], etc.

By analyzing these research works, we can notice that they have two key limitations: the first one is the cyber-physical interoperability problem of digital twins. In fact, all these works propose approaches for maintenance optimization using digital twins, but they do not mention how to establish the connection between the two physical and virtual spaces of the digital twin, which is considered as a major problem to overcome. The second limitation is that none of these proposed approaches address or automate all the essential points of the maintenance process at once. It is necessary that the proposed approach be global and treat these essential points, in particular: automatic data management, real-time failure detection, prediction of future failures, automation of the maintenance policy choice, optimization of corrective and preventive maintenance, establishment of maintenance plans, automation of decision making etc.

III. The Global Proposed MMSDTO Methodology

The principal goal of this paper is to automate and optimize the maintenance management of manufacturing processes. Therefore, a structured methodology based on digital twins (DT) and ontologies is proposed. In fact, these two concepts are merged and integrated in the field of maintenance, which gives birth to MMSDTO (Maintenance Management System based on Digital Twins and Ontologies) methodology.

It is decomposed into two main phases as indicated in Fig. 1, namely: Construction phase and operation phase.

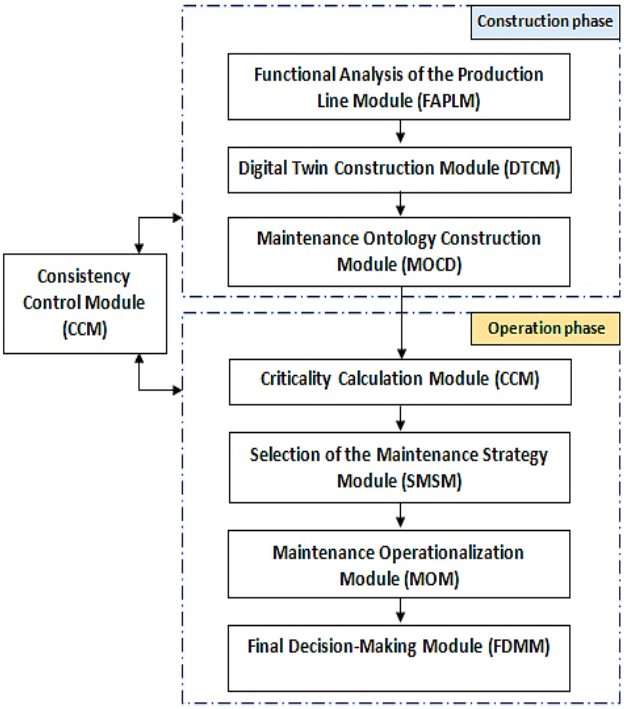


Fig. 1. The Global Proposed MMSDTO Methodology.

First, the construction phase contains three modules:

- Functional Analysis of the Production Line Module (FAPLM): the good knowledge of the production system is an essential factor to have a successful study, for this reason this first module of the construction phase represents one of the pillars of the MMSDTO methodology. Its usefulness lies in the fact that it collects all the necessary information (blocks, components and sub-components of the production lines, production parameters, etc.).
- Digital Twin construction module (DTCM): In this module, the Digital Twin will be constructed with consideration of its standard structure [9]. In fact, a virtual counterpart will be established from the physical entity studied. Moreover, the cyber-physical connection between these two spaces, serving for the transfer of the collected data, the deduced information and the necessary services for the functioning of the system, will be realized.

- Maintenance Ontology Construction Module (MOCD): Through this last module of the construction phase, the expressiveness capacity that ontologies possess will be exploited to build a maintenance ontology (DTM-Onto). The available data will be expressed in a standard language, which will allow converting them to a semantic model that will be used to achieve interoperability between the different elements of the DT.

The operation phase contains four modules:

- Criticality Calculation Module (CCM): The DTM-Onto previously constructed will be enriched by rules of computation and classification of the criticalities of the different elements of the studied system. This will be used to highlight the critical elements of the system to which priority will be given in the maintenance programs.
- Selection of the Maintenance Strategy Module (SMSM): After the analysis of the failure modes of each element as well as the evaluation of the criticality associated to each mode in the previous module, a hierarchization of the different criticality indexes is done in this module, in order to choose the adequate maintenance policy.
- Maintenance Operationalization Module (MOM): The DTMa-Onto will be enriched by other calculation rules to generate the various results and information necessary for the establishment of maintenance plans and planning.
- Final Decision-Making Module (FDMM): Based on the results obtained by the ontologies, maintenance actions must be planned. Thus, maintenance plans and planning will be realized to synthesize the work realized.
- Consistency Control Module (CCM): This module plays a key role in maintaining consistency between the modules of the two main phases of the global methodology.

These modules are executed according to a working process, as shown in Fig. 2.

The first step of the operating system of the proposed MMSDTO methodology is the functional decomposition of the studied machine in several blocks, then in several elements. This will allow to build the Digital Twin of the machine and to feed the ontology with the necessary data and inference rules. In fact, the ontology will help to solve the cyber-physical interoperability problem between the elements of the Digital Twin. In addition, it will generate several computation and classification results: criticalities of the machine elements, TBF, MTBF, Weibull parameters, etc. Finally, all these results will lead to the establishment of global maintenance plans and planning for the production equipment.

All the modules of the methodology will be detailed in the following sections.

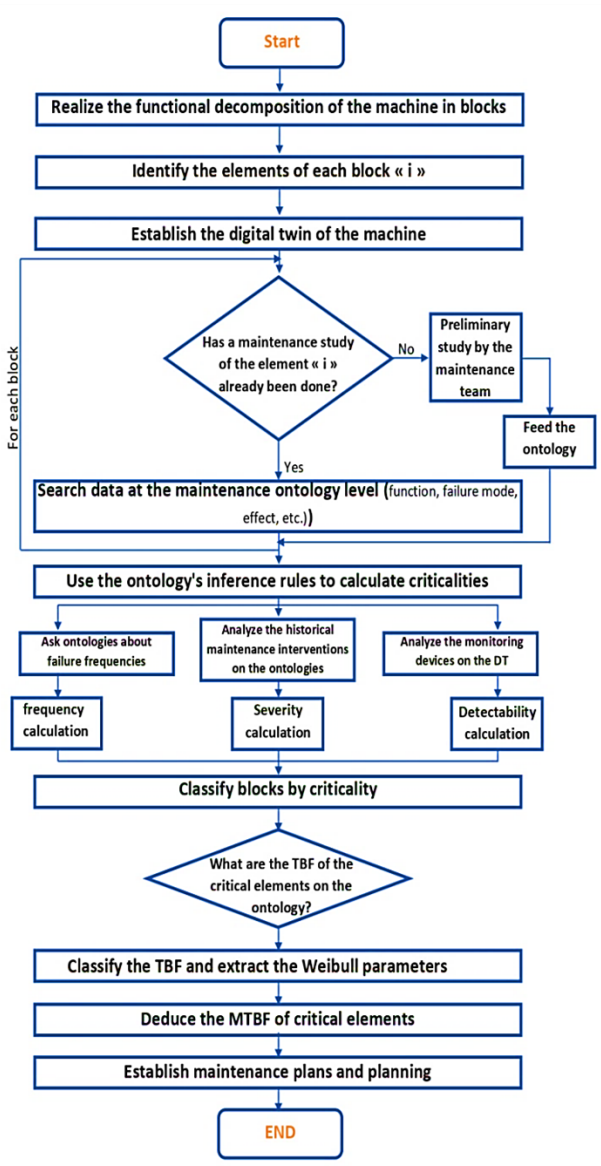


Fig. 2. The Working Process of the MMSDTO System.

IV. Functional Analysis of the Production Line Module

This module consists in identifying the components of the production system studied as well as the functions they must ensure. Therefore, it is a question of identifying the functions performed by the system and the technological solutions that achieve these functions. To do this, the following elementary phases are implemented:

- A decomposition of the system into blocks, components and sub-components in an exhaustive or limited way, up to the desired level of decomposition according to the needs and the objectives of the study.
- A description of the functions performed by each element.
- A preliminary identification of the dependencies (or cause & effect) between these functions.

V. Digital Twin Construction Module

The main objective of this module is the construction of the Digital Twin which requires the validation of the five elements of its standard structure, namely: Physical entity, virtual model, connection model, services and DT data model: [9].

Firstly, a passage from the functional decomposition of the production line module is crucial, at the level of which the latter is decomposed into several blocks, the blocks into components and the components into sub-components. In fact, this decomposition serves to simplify the virtual reproduction of the functioning of the physical entities, as well as to identify the zones that must be reinforced by sensors to achieve a perfect similarity between the physical and virtual spaces, to increase the detection of failures and to ensure a good follow-up of the production and maintenance. On the other hand, on the virtual level, the components of the manufacturing system are represented by geometrical models in CAD software. In addition, the flows (production flows, logistic flows, etc.) and the behavioral models (fatigue, elasticity, etc.) are simulated respectively on flow and behavioral modeling software.

At this stage, and to ensure a faithful exchange of data and services between the two spaces, a cyber-physical connection must be established. Normally, this connection is achieved using artificial intelligence tools and monitoring information systems [23] [24], but this still has some shortcomings in the industrial context, namely: the difficulty of preserving the semantics of the transmitted data, the inability to transfer a considerable quantity of information between the different actors of the system and the difficulty of logical reasoning. So, to overcome these problems, the concept of ontologies is integrated. This concept has already been integrated in one of our previous paper [23], through the construction of a production ontology that solves the problem of interoperability between physical entities and their virtual counterparts. In this paper, another maintenance ontology (DTMa-Onto) is added. The DTMa-Onto will be enriched with inference rules (maintenance rules, prediction rules, optimization rules, etc.), in order to be able to establish, at the end, the maintenance plans and planning.

In fact, both ontologies will be fed with the necessary data. This data will be stored in the ontologies, processed by the ontology inference processors, and transferred to the virtual space. Afterwards, new information can be generated.

This information will also be stored and processed by the ontologies. The cycle repeats itself in order to control and master the production and maintenance.

The operating system of the Digital Twin is clearly schematized in Fig. 3.

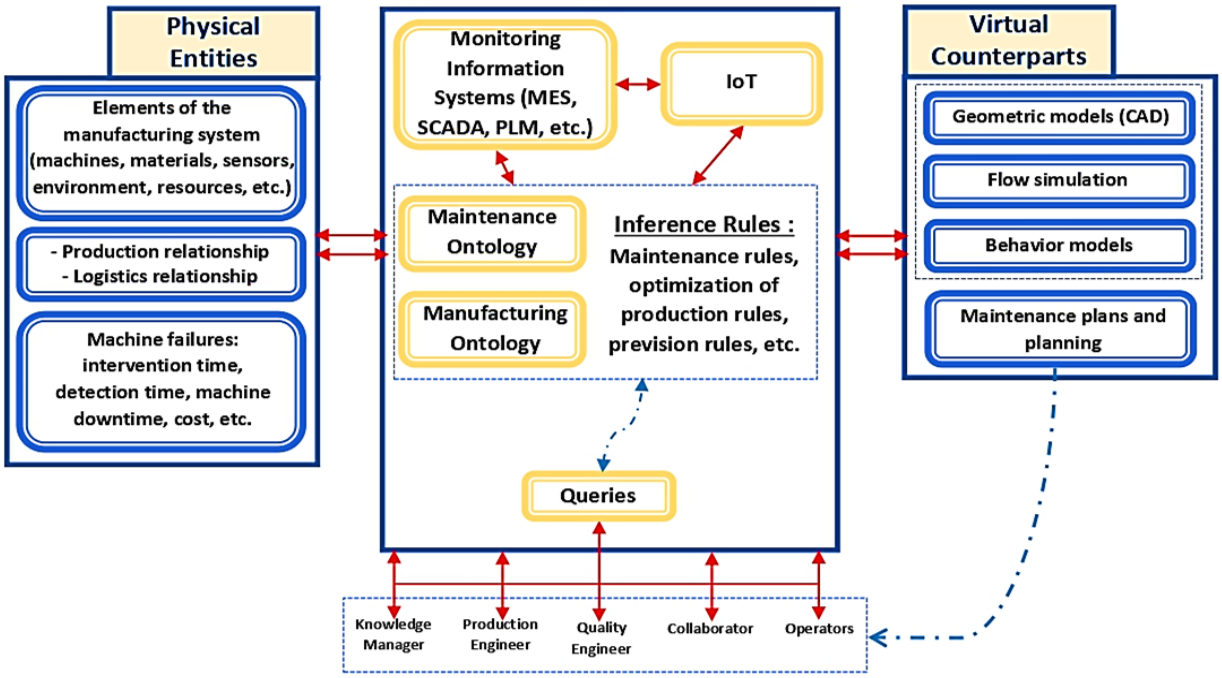


Fig. 3. The Operating System of the Digital Twin is Clearly Schematized in Figure 3.

VI. Maintenance Ontology Construction Module

This module represents the starting point for the implementation of the later phases.

At this level, a maintenance ontology (DTMa-Onto) will be constructed, as shown in Fig. 4, which will solve a large part of the problem of interoperability between the two physical and virtual spaces of the Digital Twin, due to its capacity to exchange a large quantity of data between people and/or machines, to analyze the information exchanged and to reuse it [25].

This will help to reproduce the operation of the production system digitally, as well as detect anomalies and report failures. In fact, the construction of the DTMa-Onto recognizes that it must pass through several main stages and that are realized on the editor of the ontologies Protégé.

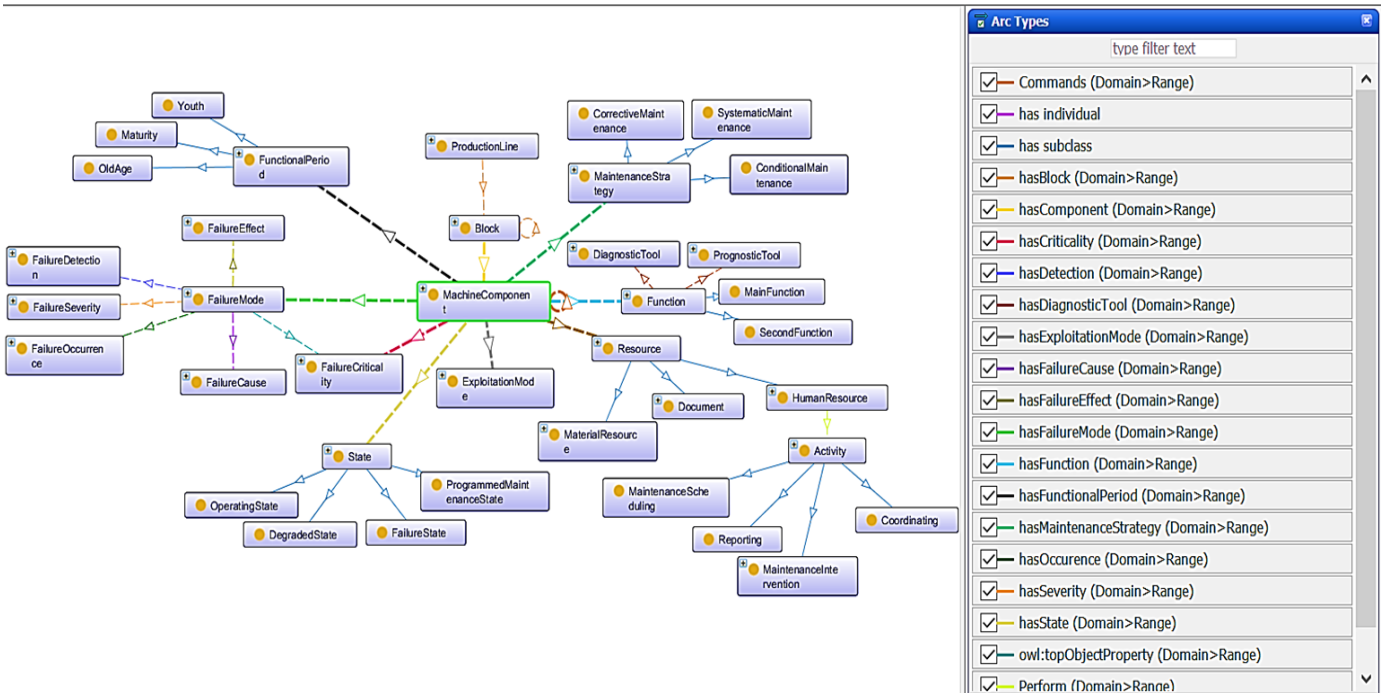


Fig. 4. The Constructed DTMa-Onto.

A. Class Definition

This phase consists in the determination of the ensemble of individuals in a specific domain. The domain considered in our case is the maintenance domain. In this way, these classes are divided into three categories:

- Classes related to the production line: they contain the different blocks, components and subcomponents of the machine, the operating mode as well as the human and material resources necessary for the operation of the production line, etc.
- Classes related to the maintenance of the production line: they comprise the failure modes, the human and material resources necessary for diagnosis and maintenance, the operating period, etc.
- Classes related to the choice of maintenance strategies: they include the different types of possible maintenance (corrective maintenance, preventive maintenance, etc.).

The whole set of classes is represented in Fig. 5.

B. Object and Data Properties Definition

First, this step consists of defining the object properties. In other words, the relationships between classes and individuals must be determined, while defining their domains, ranges and inverse properties.

Secondly, we need to specify the data properties. These relate a predicate to a single subject to an attribute data form, which can be a string, an integer, a real number, a date, etc. Fig. 6 recapitulates all the object properties and the data properties of the DTMa-Onto.

It should be noted that the implementation of all these steps leads to a hierarchy of the classes and subclasses of the production line, the maintenance of its different parts and the choice of the appropriate maintenance strategy for each situation.

The particularity of the proposed DTMa-Onto resides in the fact that it is standard and can be adapted to the maintenance of any industrial manufacturing process.

In order to evaluate the consistency of the developed ontology and to execute forward the reasoning rules, we have used the description logic reasoner Pellet [25] which is integrated in the open-source ontology editor Protégé 5.0.

In the next section, the expressiveness of the OWL ontology DTMa-Onto will be enhanced with three different categories of reasoning rules modeled with SWRL Language (Semantic Web Rule Language).

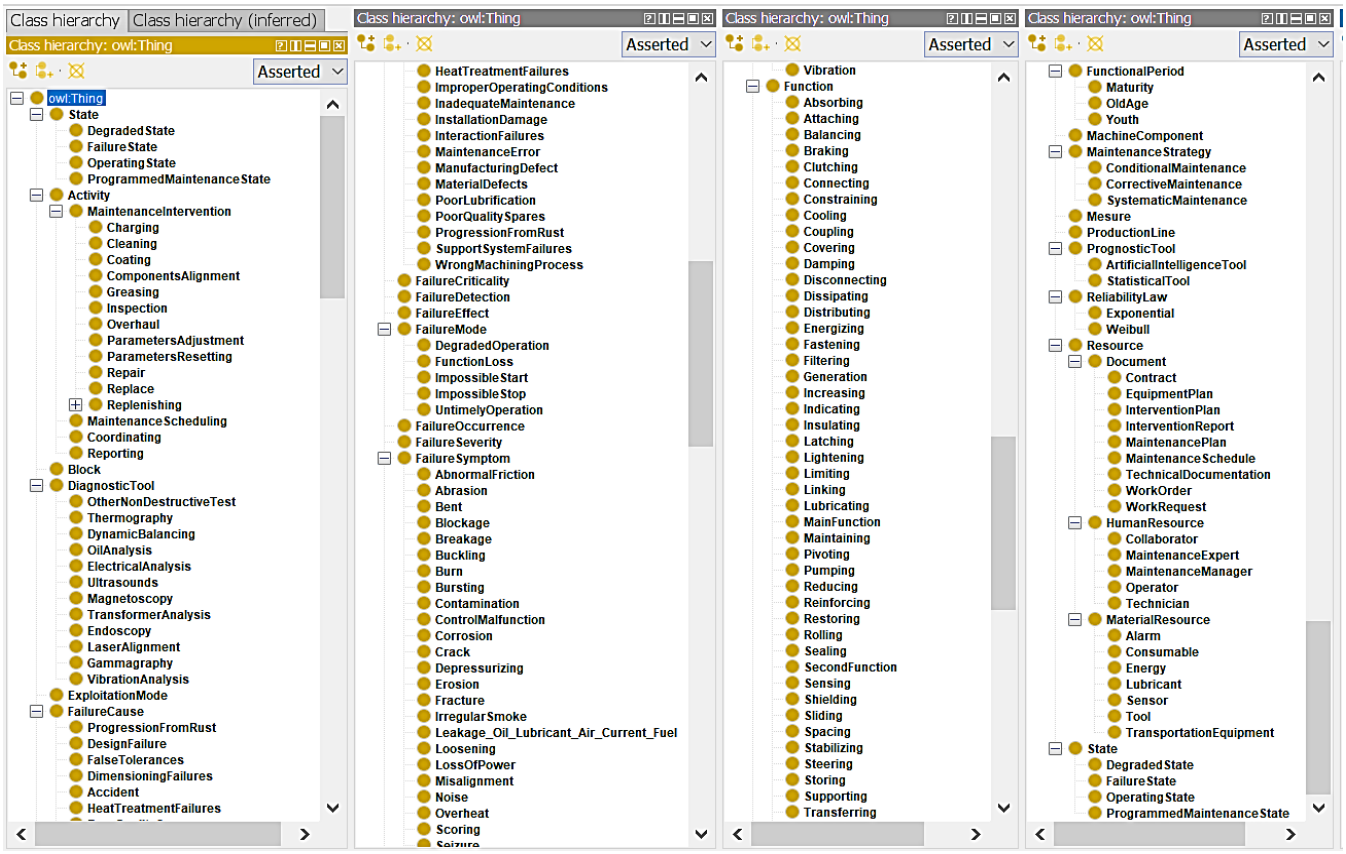


Fig. 5. All Classes of the DTMa-Onto.

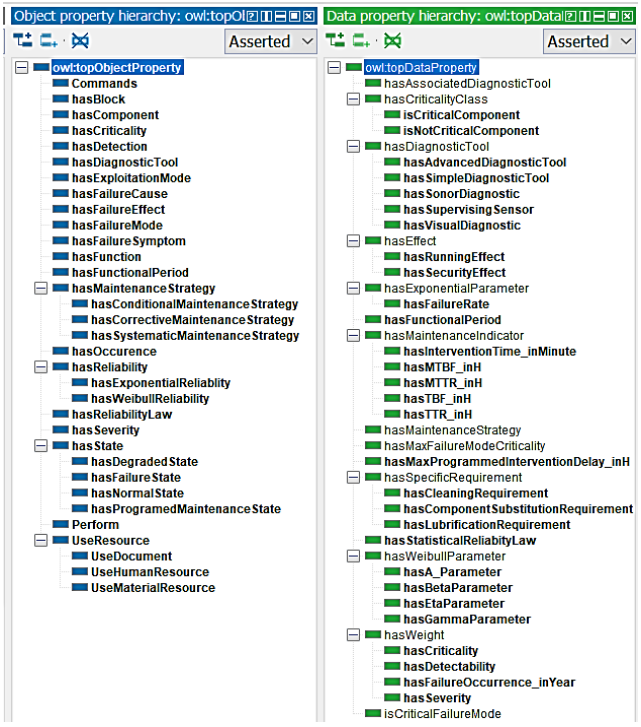


Fig. 6. Object Properties and Data Properties of the DTMa-Onto.

VII. Criticality Calculation Module

After constructing the backbone of the DTMa-Onto through the definition of classes, object properties and data properties, now it needs to be enriched with inference rules in order to perform the computation and the logical reasoning. Therefore, a first category of rules will be executed in this module, namely the rules for calculating and classifying the criticality of components. First of all, the criticality evaluation is based on the calculation and estimation of:

- The Frequency of Occurrence Index (O) (Rules R1-R4): it represents the probability that the cause of failure appears and generates the failure mode considered.
- The Severity Index (S) (Rules S1-S4): it quantifies the severity of the consequences that the failure generates.
- The Non-Detection index (D) (Rules D1-D6): it represents the probability of non-detection of the failure mode.

The criticality index (Rule C1) is calculated by multiplying the three elementary indices for each component:

$$C = O * S * D \quad (1)$$

In addition to that, the evaluation is done taking into account the current or expected state of the system, which allows to prioritize the failure modes and to identify the most critical ones to study in priority.

In addition to that, the evaluation is done taking into account the current or expected state of the system, which allows to prioritize the failure modes and to identify the most critical ones to study in priority.

As Table I shows, the calculation of these indices is formalized in the form of several rules expressed using the SWRL language.

TABLE I. RULES OF CRITICALITIES CALCULATION AND CLASSIFICATION EXPRESSED IN SWRL LANGUAGE

Criticality Rules
R1 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^hasFailureOccurrence_inYear(?M, ?o) ^ swrlb:greaterThanOrEqual(?o, 12) -> hasOccurrence(?M, 4)
R2 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^hasFailureOccurrence_inYear(?M, ?o) ^ swrlb:greaterThanOrEqual(?o, 4) ^ swrlb:lessThanorEqual(?o, 12) -> hasOccurrence(?M, 3)
R3 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^hasFailureOccurrence_inYear (?M, ?o) ^ swrlb:greaterThan(?o, 1) ^ swrlb:lessThanorEqual(?o, 4) -> hasOccurrence(?M, 2)
R4 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^hasFailureOccurrence_inYear(?M, ?o) ^ swrlb:lessThanOrEqual(?o, 1) -> hasOccurrence(?M, 1)
S1 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^hasInterventionTime_inMinute(?M, ?IT) ^ swrlb:greaterThan(?IT, 60) -> hasSeverity(?M, 4)
S2 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^hasInterventionTime_inMinute(?M, ?IT) ^swrlb:greaterThan(?IT, 20) ^ swrlb:lessThan(?IT,60) -> hasSeverity(?M, 3)
S3 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^hasInterventionTime_inMinute(?M, ?IT) ^swrlb:greaterThan(?IT, 5) ^ swrlb:lessThan(?IT, 20) -> hasSeverity(?M, 2)
S4 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^hasInterventionTime_inMinute(?M, ?IT)^swrlb:lessThan(?IT,5) -> hasSeverity(?M, 1)
D1 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^ hasSupervisingSensor(?M, 0) ^ hasVisualDiagnostic(?M, 0)^hasSonorDiagnostic(?M,0)^hasSimpleDiagnosticTool(?M,0) ^hasAdvancedDiagnosticTool(?M,0) -> hasDetectability(?M, 4)
D2 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^ hasSupervisingSensor(?M, 0) ^ hasVisualDiagnostic(?M, 0)^hasSonorDiagnostic(?M,0)^hasSimpleDiagnosticTool(?M,0) ^hasAdvancedDiagnosticTool(?M,1) -> hasDetectability(?M, 4)
D3 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^ hasSupervisingSensor(?M, 0) ^ hasVisualDiagnostic(?M, 0)^hasSonorDiagnostic(?M,0)^hasSimpleDiagnosticTool(?M,1) ^hasAdvancedDiagnosticTool(?M,0) -> hasDetectability(?M, 3)
D4 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^ hasSupervisingSensor(?M, 0) ^ hasVisualDiagnostic(?M, 0)^hasSonorDiagnostic(?M,1)^hasSimpleDiagnosticTool(?M,0) ^hasAdvancedDiagnosticTool(?M,0) -> hasDetectability(?M, 2)
D5 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^ hasSupervisingSensor(?M, 0) ^ hasVisualDiagnostic(?M, 1)^hasSonorDiagnostic(?M,0)^hasSimpleDiagnosticTool(?M,0) ^hasAdvancedDiagnosticTool(?M,0) -> hasDetectability(?M, 1)
D6 : MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^ hasSupervisingSensor(?M, 1) ^ hasVisualDiagnostic(?M, 0)^hasSonorDiagnostic(?M,0)^hasSimpleDiagnosticTool(?M,0) ^hasAdvancedDiagnosticTool(?M,0) -> hasDetectability(?M, 1)
C1: MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^ hasOccurrence(?M, ?o) ^ hasSeverity(?M, ?s) ^ hasDetectability(?M, ?d) ^ swrlb:multiply(?q, ?o, ?s, ?d) -> hasCriticality(?M, ?q)
C2: MachineComponent(?C) ^ hasMaxFailureModeCriticality(?C, ?q) ^ swrlb:greaterThanOrEqual(?q, 32) -> hasCriticalityClass(?C, "Critical")
C3: MachineComponent(?C) ^ hasMaxFailureModeCriticality(?C, ?q) ^ swrlb:lessThan(?q, 32) -> hasCriticalityClass(?C, "NotCritical")
Q1: MachineComponent(?C) ^ hasFailureMode(?C, ?M) ^ hasCriticality(?M, ?q) ^ hasCriticalityClass (?C, ?s) -> sqwrl:select (?C , ?M , ?q, ?s) ^ sqwrl:columnNames ("MachineComponent", "FailureMode", "Failure Mode Criticality", "Component Criticality Class")
Q2: MachineComponent(?C) ^ hasCriticalityClass(?C, ?s) -> sqwrl:select(?C, ?s) ^ sqwrl:columnNames("MachineComponent", "Component Criticality Class")

In fact, the Digital Twin collects the necessary data (i.e. downtime, failures appeared, failure frequency, etc.), stores them on the DTMa-Onto. In its turn, the DTMa-Onto treats this collected information, and assigns for each index an adequate value according to a rating scale (from 1 to 4) programmed by the inference rules, and then the criticality of each component is calculated. It should be noted that the values as well as the rating criteria can change from one production process to another, but the principle remains the same.

A deployment of the results of the computation is performed. The final objective is to define and launch all the necessary actions, both corrective and preventive, taking into account the priorities highlighted by the evaluation of the failures' criticality. Indeed, we proceed to a prioritization of all the failure modes according to their criticality indexes (The rules: C2, C3, Q1, Q2). The critical points of the equipment are then identified. They correspond to the failures which have a criticality higher than a threshold predefined by the maintenance team and which takes into account the expected reliability objectives as well as the studied technologies. Priority actions must also be considered for any severity or occurrence index score equal to 4, because, even if the criticality of these failures is lower than the pre-established threshold, they represent a real risk. This constraint has been taken into account in the rules.

VIII. SELECTION OF THE MAINTENANCE STRATEGY MODULE

Fig. 7 summarizes the methodology for choosing the maintenance policy proposed.

The inference rules implemented in the criticality calculation module have led to the establishment of other rules concerning the choice of the maintenance policy.

In fact, to make the analysis of the failure modes of each component and the evaluation of the criticality associated with each mode useful, it is necessary to define for each component the type of maintenance that is appropriate. To do this, the maintenance team determines a minimum criticality threshold above which the failure modes become critical, and then the corrective maintenance is replaced by the preventive one.

However, in some cases, it is difficult to control the equipment even if the criticality index exceeds the threshold set by the working group, and therefore, its maintenance remains corrective. In addition to this, the safety of the personnel and the equipment must be taken into account.

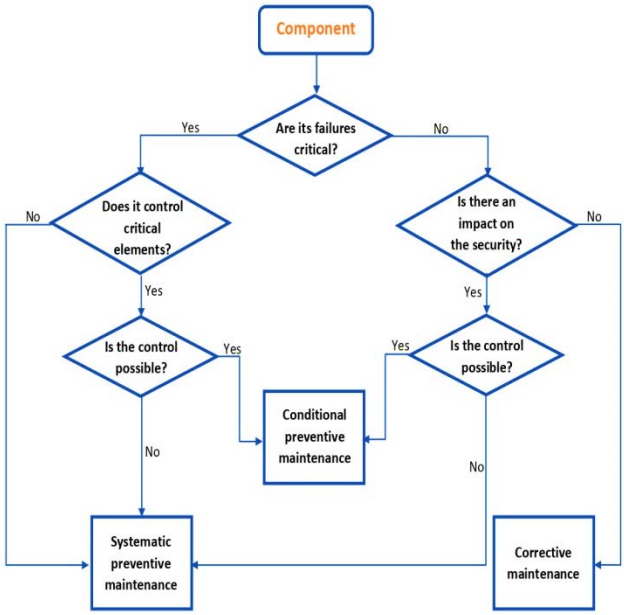


Fig. 7. The Proposed Methodology for the Maintenance Policy Selection.

This approach for selecting the maintenance strategy is programmed at the DTMa-Onto level using various rules expressed by the SWRL language. Table II groups all these rules.

TABLE II. MAINTENANCE STRATEGY RULES EXPRESSED IN SWRL LANGUAGE

Maintenance Strategy Rules
P1: MachineComponent (?C) ^ hasCriticalityClass(?C, "Critical") ^ Commands(?C, NoneComponent) -> hasMaintenanceStrategy(?C, "SystematicMaintenance")
P2: MachineComponent (?C) ^ MachineComponent (?N) ^ hasCriticalityClass(?C, "Critical") ^ Commands(?C, ?N) ^ hasCriticalityClass(?N, "NotCritical") ^ hasAssociatedDiagnosticTool(?C, 0) -> hasMaintenanceStrategy(?C, "SystematicMaintenance")
P3 : MachineComponent (?C) ^ MachineComponent (?N) ^ hasCriticalityClass(?C, "Critical") ^ Commands(?C, ?N) ^ hasCriticalityClass(?N, "NotCritical") ^ hasAssociatedDiagnosticTool(?C, 1) -> hasMaintenanceStrategy(?C, "ConditionalMaintenance")
P4: MachineComponent (?C) ^ hasCriticalityClass(?C, "NotCritical") ^ hasSecurityEffect (?C, 1) ^ hasAssociatedDiagnosticTool(?C, 1) -> hasMaintenanceStrategy(?C, "ConditionalMaintenance")
P5: MachineComponent (?C) ^ hasCriticalityClass(?C, "NotCritical") ^ hasSecurityEffect (?C, 1) ^ hasAssociatedDiagnosticTool(?C, 0) -> hasMaintenanceStrategy(?C, "SystematicMaintenance")
P6: MachineComponent (?C) ^ hasCriticalityClass(?C, "NotCritical") ^ hasSecurityEffect (?C, 0) -> hasMaintenanceStrategy(?C, "CorrectiveMaintenance")

IX. MAINTENANCE OPERATIONALIZATION MODULE

As presented in Table III, this module covers the third class of inference rules of the DTMa-Onto.

These rules are a formalization of the two statistical laws of component reliability, namely:

- The exponential law: where the failure rate is constant. It is valid in the case of electrical components and components in maturity phase.

$$R(t) = exp(-\lambda t) \quad (2)$$

With:

✓ λ is the failure rate. It represents the proportion of defective parts that we obtain during a very short time interval

- The Weibull law: It is valid in the general case and takes into account the three phases of the life cycle (Growth, maturity and decline)

$$R(t) = \exp(-(\frac{t-\gamma}{\eta})^{\beta} \quad (3)$$

With:

✓ β is the shape parameter

✓ η is the scale parameter

✓ γ is the position parameter

TABLE III. MAINTENANCE OPERATIONALIZATION RULES EXPRESSED IN SWRL LANGUAGE

Maintenance Operationalization Rules
W1: MachineComponent(?C) ^ ReliabilityLaw(?R) ^ hasReliabilityLaw (?C, ?R) ^ hasStatisticalReliabityLaw(?R, "Weibull") ^ hasBetaParameter(?R, ?b) ^ hasGammaParameter(?R, ?g) ^ hasEtaParameter(?R, ?e) ^ hasA_Parameter(?R, ?A) ^ swrlb:multiply(?t, ?A, ?e) ^ swrlb:add(?u, ?t, ?g) -> hasMTBF_inH(?C, ?u) ^ hasMaxProgrammedInterventionDelay_inH(?C, ?u)
W2: MachineComponent(?C) ^ hasFunctionalPeriod(?C, "Maturity") -> hasStatisticalReliabilityLaw(?R, "Exponential")
W3: MachineComponent(?C) ^ hasReliabilityLaw(?C, ?R) ^ hasStatisticalReliabilityLaw(?R, "Exponential") ^ hasFailureRate(?R, ?l) ^ swrlb:divide(?u, 1, ?l) -> hasMTBF_inH(?C, ?u) ^ hasMaxProgrammedInterventionDelay_inH(?C, ?u)

This module works in interaction with the numerical analysis environment and the programming language "Matlab". Indeed, using the TBF (Time between failures) extracted from the failure histories collected by the Digital Twin and stored in the DTMa-Onto, the approach developed by [26] will be applied for the estimation of the exponential and Weibull law parameters of the concerned components. These parameters will constitute the inputs for the SWRL rules of the ontology. These rules model the exponential (rules W2-W3) and Weibull (rules W1) reliability laws, and enable the calculation of MTBF (Mean time between failures) and the prediction of the next component failures, which will allow to schedule future maintenance interventions before the machine components fail. The MTBF formula is as follows:

$$R(t) = \exp(-(\frac{t-\gamma}{\eta})^{\beta}) \quad (4)$$

With:

- ✓ β is the shape parameter
- ✓ η is the scale parameter
- ✓ γ is the position parameter

X. Final Decision-Making Module

From the maintenance Operationalization module, the TBFs and MTBFs of the different critical components are calculated, and the life phase of each element is determined. Based on these calculation results obtained by the DTMa-Onto, the appropriate maintenance actions (lubrication, greasing, etc.) must be planned before the end of the TBFs. All this work gives birth to the maintenance plans and planning, which will be stored and archived at the DTMa-Onto level, in order to benefit from the experience feedback for future maintenance situations. This phase has two objectives, namely the identification and standardization of good practices and methods, as well as the transmission to the design all the experience acquired (means, processes, operating modes, etc.). The obvious goals are to capitalize successful actions and generalize them.

XI. Case of Study

The objective of this part is to validate the applicability of the proposed methodology MMSDTO, as well as its effectiveness for the resolution of maintenance problems of industrial production systems. This validation will be done through a concrete example of a company that operates in the agro alimentary industry. In fact, this company works on dairy products manufacturing (milk, yogurt, etc.), but the case study will only focus on the yogurt manufacturing and packaging process.

The first step of the process is the unrolling of the plastic, which is done by two rollers driven by a gear motor. The plastic roll is introduced in the insertion section, and in order to ensure its pecking, it passes directly into a heating system to heat its edges, due to heating resistances and temperature probes. This operation is done by a chain and serves to transmit the plastic strip during the rest of the process. Before moving to the forming block, the plastic strip passes through two ionizing dust collectors. The first one ionizes the plastic strip to prevent electrostatic sticking and to remove any foreign body. The second is a tunnel of ultraviolet lamps, which ionizes the operculum to prevent electrostatic sticking and to aspirate the existing particles. Once the plastic and the operculum are ionized, the product cups are formed using a mold and punches according to the desired specifications. The cups are then dosed with the products, which have already been prepared in the process section, using a piston dosing unit. Then the operculum is welded onto the dosed cups using a cam press and the product packs are cut out. After the product packs have been cut, they are transported to the case packer by a conveyor system to be packed in empty boxes. The last step of the packaging process is the manual pallet handling of the full boxes, and their storage in cold rooms.

The competition in the market is fierce and the interruption of production generates huge losses for the company (market shares, profits, etc.), that's why the control of the production equipment maintenance is a necessity. To this effect, the MMSDTO methodology is applied.

The first step in the approach is the functional decomposition of the production process into blocks, the blocks into components and the components into sub-components. This decomposition will be used, initially, to locate the areas of the production line that require the installation of sensors for the data transfer between the two physical and virtual spaces of the Digital Twin. Fig. 8 and 9 show, respectively, an example of the installation of the position determination sensors and an extract from the virtual part of the Digital Twin.

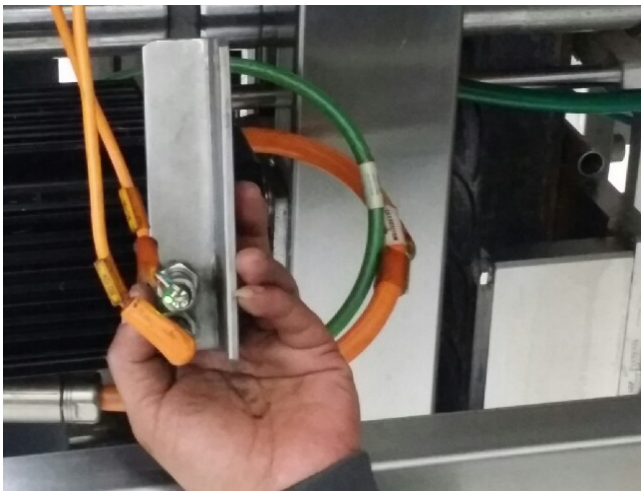


Fig. 8. Example of Sensors Installation Necessary for the Digital Twin Functioning.

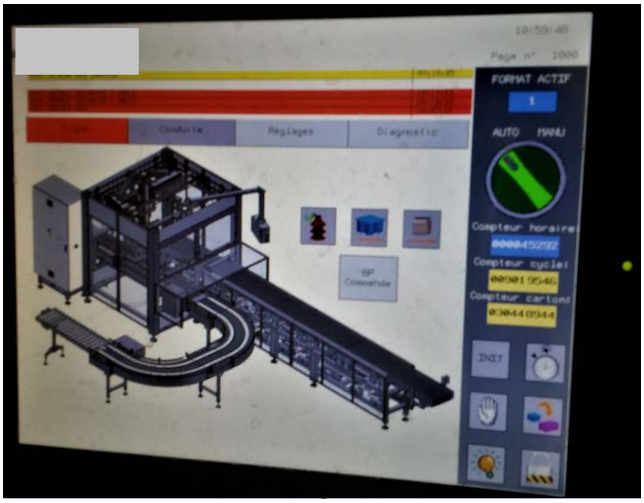


Fig. 9. An Extract from the Virtual Part of the Digital Twin.

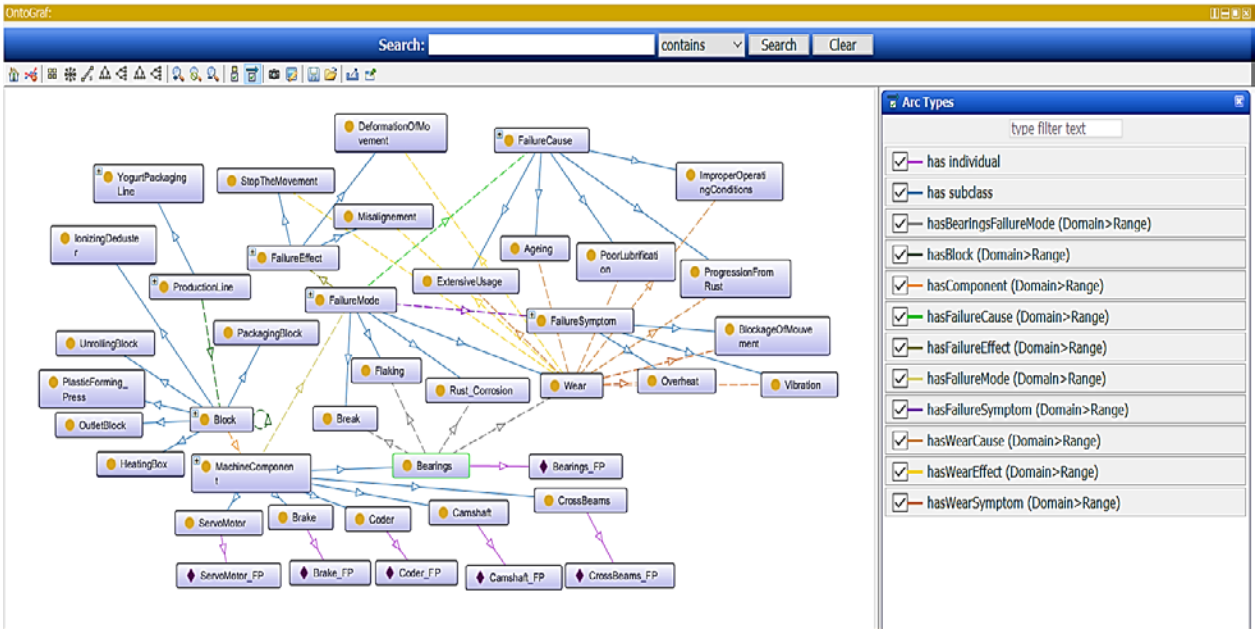


Fig. 10. The DTMa-Onto related to the Case of Study.

The second objective of this decomposition is the preliminary construction of the DTMa-Onto. Thus, the classes, the object properties and the data properties of the forming press are entered on the ontology (the work will be focused on the forming block to simplify the case of study and make it clearer). In fact, as it is represented on Fig. 10, the DTMa-Onto schematizes the different blocks of the production line, the failure modes of the different components of the forming block (the studied block), their causes, their symptoms and their effects.

Now, based on the collected data from the Digital Twin, in particular the machine downtime tracking, the DTMa-Onto, and due to its information storage and logical reasoning capabilities, assigns for each failure mode a suitable occurrence, severity and detectability value. Then, the DTMa-Onto immediately generates the results of the calculation and classification of the failure modes' criticalities, using the inference rules listed in the Table IV.

The totality of the calculation and criticality classification results of the other failure modes of the forming press components is presented in the Fig. 12.

After the classification of the failure modes by their criticality, the ontology lists all the critical elements based on the criticality threshold set by the maintenance team.

An extract of the obtained results are illustrated in the Fig. 11. These results concern the criticalities of the forming press bearings failure modes.

At this stage, the second category of inference rules, namely the rules for the maintenance policy choice, is applied. Thus, the type of maintenance for each component is determined by the DTMa-Onto.

The third category of rules, including maintenance operationalization rules, will be applied on the brake and bearings of the forming press to calculate the proper MTBFs. Fig. 14 shows the results obtained.

The Fig. 13 summarizes the results of the appropriate maintenance policy for each element. The results obtained indicate that the bearings and the brake of the forming press require a systematic preventive maintenance policy.

Finally, from all the results obtained before, it only remains to establish the general maintenance plans and planning, in order to control the maintenance process of the production line, and subsequently improve its performance.

It is necessary to note that the maintenance plans and planning are not presented in order to avoid burdening the study case with too much information and results.

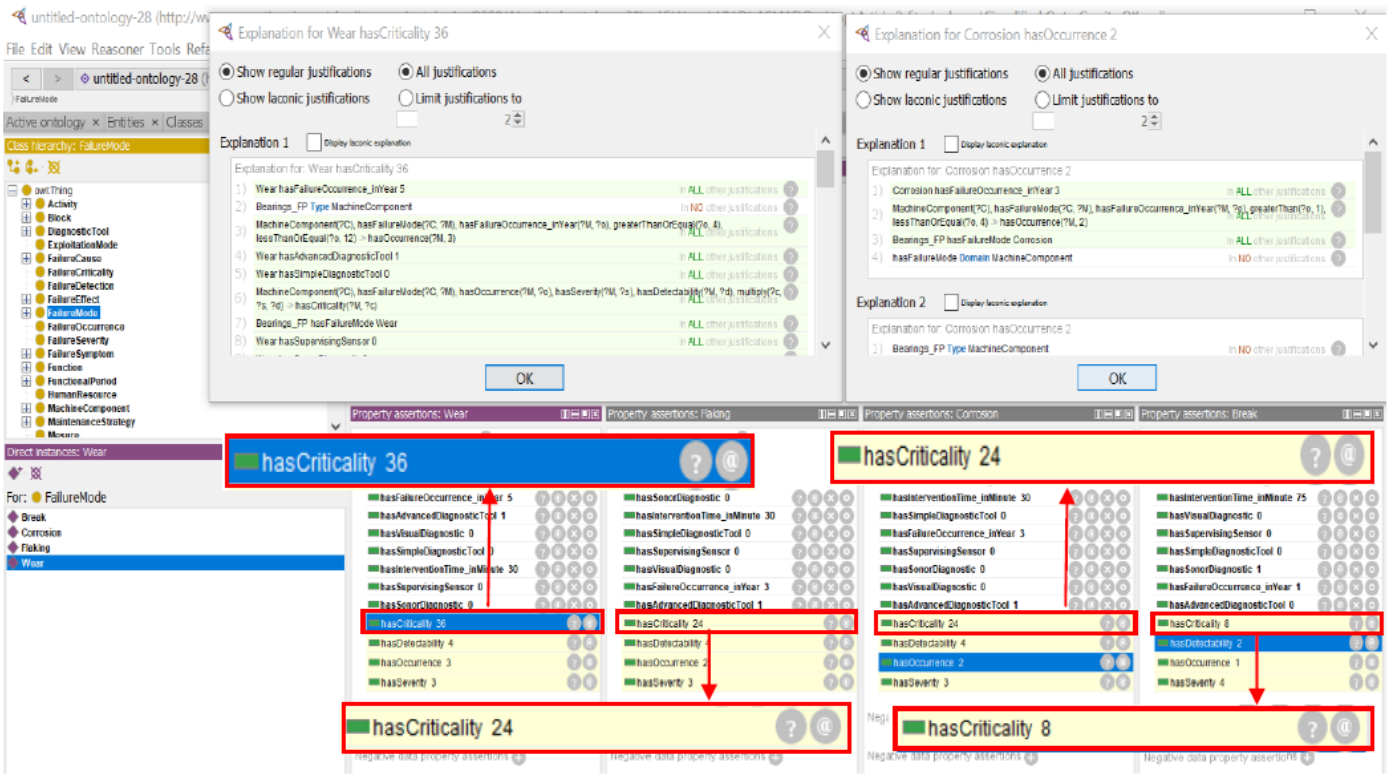
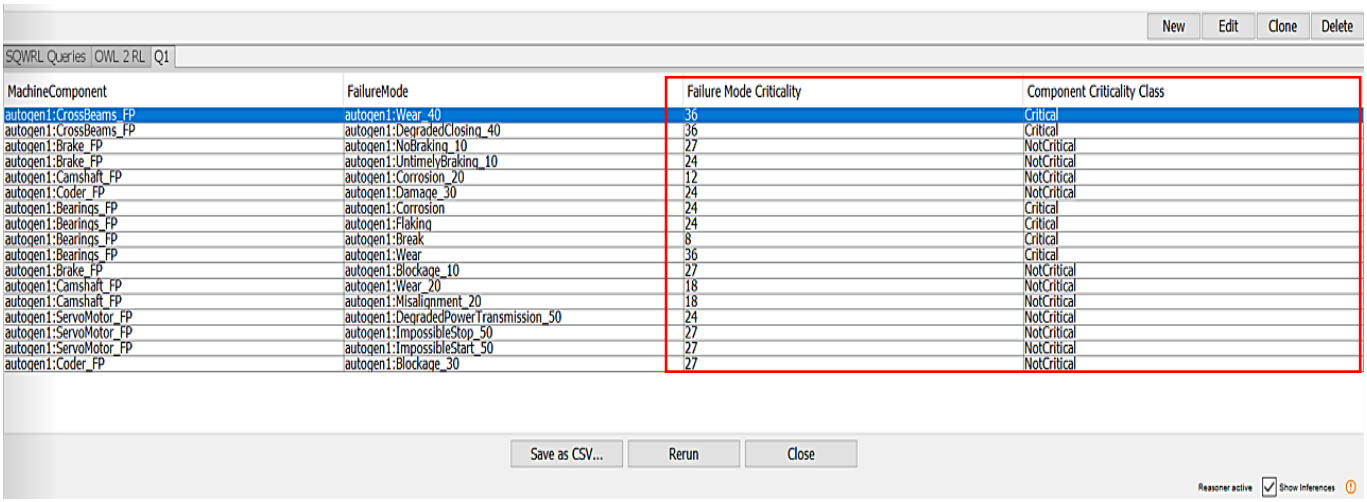


Fig. 11. An Extract of the Criticalities Calculation Results Obtained.



MachineComponent	FailureMode	Failure Mode Criticality	Component Criticality Class
autogen1:CrossBeams_FP	autogen1:Wear_40	36	Critical
autogen1:CrossBeams_FP	autogen1:DegradedClosing_40	36	Critical
autogen1:Brake_FP	autogen1:NoBraking_10	27	NotCritical
autogen1:Brake_FP	autogen1:UntimelyBraking_10	24	NotCritical
autogen1:Camshaft_FP	autogen1:Corrosion_20	12	NotCritical
autogen1:Coder_FP	autogen1:Damage_30	24	NotCritical
autogen1:Bearings_FP	autogen1:Corrosion	24	Critical
autogen1:Bearings_FP	autogen1:Flaking	24	Critical
autogen1:Bearings_FP	autogen1:Break	8	Critical
autogen1:Bearings_FP	autogen1:Wear	36	Critical
autogen1:Brake_FP	autogen1:Blockage_10	27	NotCritical
autogen1:Camshaft_FP	autogen1:Wear_20	18	NotCritical
autogen1:Camshaft_FP	autogen1:Misalignment_20	18	NotCritical
autogen1:ServoMotor_FP	autogen1:DegradedPowerTransmission_50	24	NotCritical
autogen1:ServoMotor_FP	autogen1:ImpossibleStop_50	27	NotCritical
autogen1:ServoMotor_FP	autogen1:ImpossibleStart_50	27	NotCritical
autogen1:Coder_FP	autogen1:Blockage_30	27	NotCritical

Fig. 12. The Calculation and Criticality Classification Results of the other Failure Modes of the Forming Press Components.

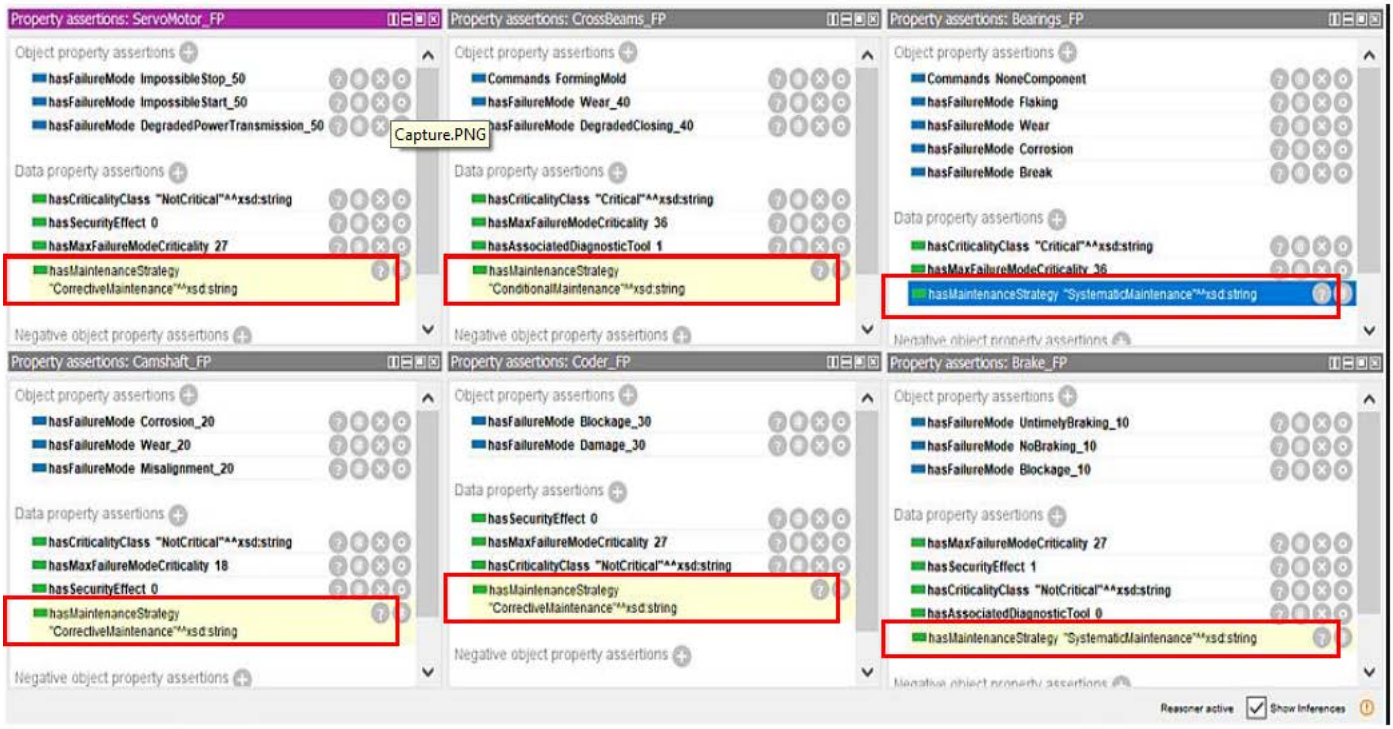


Fig. 13. The Results of the Appropriate Maintenance Policy for each Element of the Forming Press.

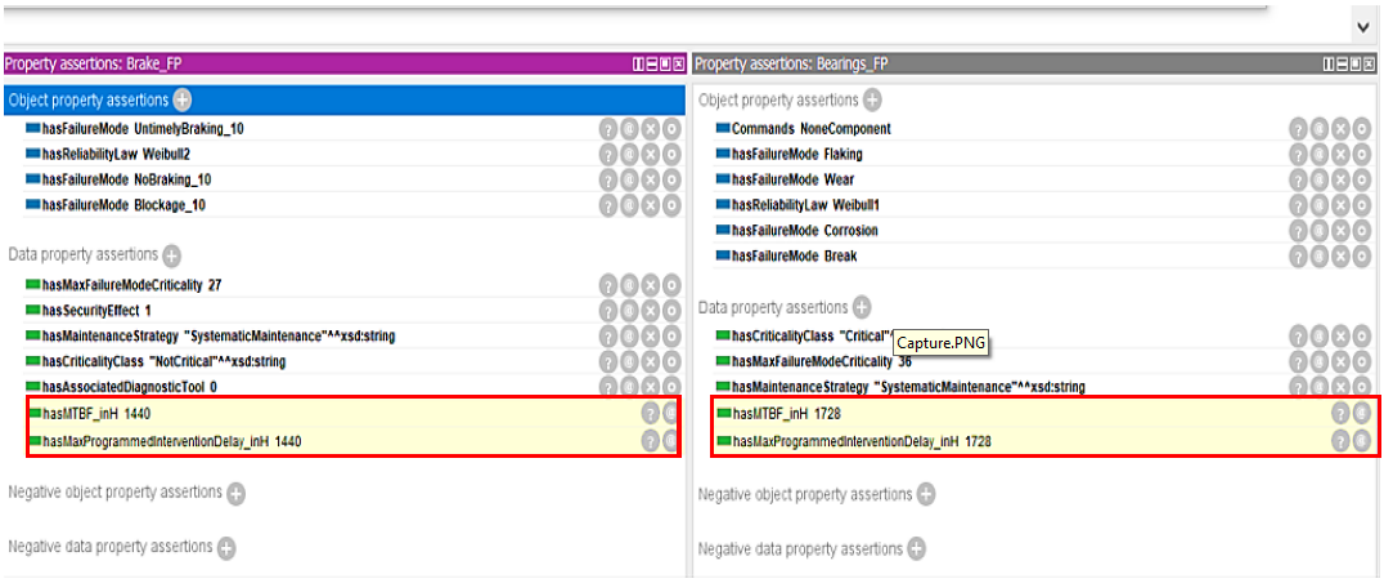


Fig. 14. The Obtained Results by the Maintenance Operationalization Module Rules.

XII. Conclusion and Perspectives

An intelligent decision making system for the automation of industrial production systems maintenance using digital twins and ontologies is developed in this paper. The integration of these two concepts in the same approach applied to the maintenance process has given birth to a hybrid system that presents new originalities.

On the first hand, the majority of research works related to digital twins propose a variety of approaches in relation to different domains and processes in general, and in relation to the maintenance process in particular, but they do not take into account the dimension of the connection between the physical and virtual spaces, which is essential for the applicability of these approaches. Thus, the resolution of this problem of cyber-physical interoperability of digital twins was the first objective to be attained through this paper. In fact, the concept of ontologies was integrated in the proposed approach. A maintenance ontology (DTMa-Onto) was constructed, and due to the expressivity capacities of ontologies, the DTMa-Onto

was employed to ensure in real time a faithful transfer of large quantities of data between the two physical and virtual spaces in order to reproduce the functioning and behavior of physical entities digitally. On the other hand, the second originality of this paper is that it proposes an automation approach for the control of the whole maintenance process, contrary to previous works which only treat some sides of the maintenance process in the same approach and cannot take the decision automatically. Indeed, the DTMa-Onto has been fed by many categories of inference rules that are used to calculate and classify criticalities and downtimes, choose the maintenance policy, propose actions to be taken, predict failures, etc. To summarize, the work realized is in the form of a structured and global methodology for the automation of the entire maintenance process, from data collection to decision making and the methodology was validated by an industrial case study.

As perspectives, it is suggested to integrate other aspects in the proposed approach (the environmental aspect, the financial aspect, etc.). It is also proposed to automate the maintenance process in another way, using other artificial intelligence tools. Another perspective is to propose other approaches for the automation of other processes (logistics, design, etc.).

REFERENCES

[1] VanDerHorn, E., & Mahadevan, S. (2021). Digital Twin: Generalization, characterization and implementation. *Decision Support Systems*, *145*, 113524.

[2] Ghita, M., Siham, B., & Hicham, M. (2020). Digital twins development architectures and deployment technologies: Moroccan use case. *International Journal of Advanced Computer Science and Applications*, *11*(2).

[3] Pang, T. Y., Pelaez Restrepo, J. D., Cheng, C. T., Yasin, A., Lim, H., & Miletic, M. (2021). Developing a digital twin and digital thread framework for an 'Industry 4.0'Shipyard. *Applied Sciences*, *11*(3), 1097.

[4] Grieves, M. (2014). Digital twin: manufacturing excellence through virtual factory replication. White paper, 1(2014), 1-7.

[5] Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In Transdisciplinary perspectives on complex systems (pp. 85-113). Springer, Cham.

[6] Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital Twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, *51*(11), 1016-1022.F

[7] Rasheed, A., San, O., & Kvamsdal, T. (2020). Digital twin: Values, challenges and enablers from a modeling perspective. *Ieee Access*, *8*, 21980-22012.

[8] Abadi, M., Abadi, C., Abadi, A., & Ben-Azza, H. (2022). Digital Twin-Driven Approach for Smart Industrial Product Design. In *International Conference On Big Data and Internet of Things* (pp. 263-273). Springer, Cham.

[9] Tao, F., Zhang, M., Liu, Y., & Nee, A. Y. (2018). Digital twin driven prognostics and health management for complex equipment. *Cirp Annals*, *67*(1), 169-172.

[10] Tao, F., & Zhang, M. (2017). Digital twin shop-floor: a new shop-floor paradigm towards smart manufacturing. *Ieee Access*, *5*, 20418-20427.

[11] Souifi, A., Cherfi, Z., Marc, Z., Barkallah, M., & Haddar, M. (2021, March). Le jumeau numérique dans le pilotage de performance. In *17ème colloque national S-mart AIP-PRIMECA*.

[12] He, B., Li, T. Y., & Xiao, J. L. (2022). Digital Twin-Driven Design for Product Control System. In *Digital Twins for Digital Transformation: Innovation in Industry* (pp. 41-65). Springer, Cham.

[13] Zhao, Z., Zhang, M., Chen, J., Qu, T., & Huang, G. Q. (2022). Digital Twin-enabled Dynamic Spatial-temporal Knowledge Graph for Production Logistics Resource Allocation. *Computers & Industrial Engineering*, 108454.

[14] Nguyen, T., Duong, Q. H., Van Nguyen, T., Zhu, Y., & Zhou, L. (2022). Knowledge mapping of digital twin and physical internet in supply chain management: A systematic literature review. *International Journal of Production Economics*, *244*, 108381.

[15] Fu, Y., Zhu, G., Zhu, M., & Xuan, F. (2022). Digital Twin for Integration of Design-Manufacturing-Maintenance: An Overview. *Chinese Journal of Mechanical Engineering*, *35*(1), 1-20.

[16] Errandonea, I., Beltrán, S., & Arrizabalaga, S. (2020). Digital Twin for maintenance: A literature review. *Computers in Industry*, *123*, 103316.

[17] Pairet, È., Ardón, P., Liu, X., Lopes, J., Hastie, H., & Lohan, K. S. (2019, March). A digital twin for human-robot interaction. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 372-372). IEEE.

[18] Longo, F., Nicoletti, L., & Padovano, A. (2019). Ubiquitous knowledge empowers the Smart Factory: The impacts of a Service-oriented Digital Twin on enterprises' performance. *Annual Reviews in Control*, *47*, 221-236.

[19] Rødseth, H., Wilson, A., & Schjølberg, P. (2020). Integrated Production and Maintenance Planning for Successful Asset Management Strategy Implementation. In *Engineering Assets and Public Infrastructures in the Age of Digitalization* (pp. 610-617). Springer, Cham.

[20] Eckhart, M., & Ekelhart, A. (2019). Digital twins for cyber-physical systems security: State of the art and outlook. *Security and quality in cyber-physical systems engineering*, 383-412.

[21] Semeraro, C., Lezoche, M., Panetto, H., & Dassisti, M. (2021). Digital twin paradigm: A systematic literature review. *Computers in Industry*, *130*, 103469.

[22] Aivaliotis, P., Georgoulias, K., & Chryssolouris, G. (2019). The use of Digital Twin for predictive maintenance in manufacturing. *International Journal of Computer Integrated Manufacturing*, *32*(11), 1067-1080.

[23] Abadi, M., Abadi, C., Abadi, A., & Ben-Azza, H. (2022). A Smart Decision Making System for the Selection of Production Parameters using Digital Twin and Ontologies. *International Journal of Advanced Computer Science and Applications*, *13*(2).

[24] Meyer, M., Yu, Z., Gulati, P., Delforouzi, A., Roggenbuck, J., & Wolf, K. Ontologies for digital twins in smart manufacturing Whitepaper.

[25] Abadi, A., Sekkat, S., & Ben-Azza, H. (2015, December). Utilisation des ontologies pour supporter la conception simultanée du produit et de sa chaîne logistique. In *Xème Conférence Internationale: Conception et Production Intégrées*.

[26] Mfetoum, I. M., Essiane, S. N., Kamanke, E. F., Ndzie, G. B., & Mbemmo, S. (2016). Estimation des paramètres du modèle de WEIBULL: application à la modélisation de la fiabilité de l'entailleuse-perceuse mécanique de ballasts de chemin de fer de CAMRAIL. *Sciences, Technologies et Développement, Edition spéciale, pp117-121*.

An Improved Arabic Sentiment Analysis Approach using Optimized Multinomial Naïve Bayes Classifier

Ahmed Alsanad
STC's Artificial Intelligence Chair, Department of Information Systems
College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

***Abstract*—Arabic sentiment analysis has emerged during the last decade as a computational process on Arabic texts for extracting people's attitudes toward targeted objects or their feelings and emotions regarding targeted events. Sentiment analysis (SA) using machine learning (ML) methods has become an important research task for developing various text-based applications. Among different ML classifiers, multinomial Naïve Bayes (MNNB) classifier is widely used for documents classification due to its ability for performing statistical analysis of text contents. It significantly simplifies textual-data classification and offers an alternative to heavy ML-based semantic analysis methods. However, the MNNB classifier has a number of hyper-parameters affects the classification task of texts and controls the decision boundary of the model itself. In this paper, an optimized MNNB classifier-based approach is proposed for improving Arabic sentiment analysis. A number of experiments are conducted on large sets of Arabic tweets to evaluate the proposed approach. The optimized MNNB classifier is trained on three datasets and tested on a different separated test set to show the performance of developed approach. The experimental results on the test set revealed that the optimized MNNB classifier of proposed approach outperforms the traditional MNNB classifier and other baseline classifiers. The accuracy rate of the optimization approach is increased by 1.6% compared with using the default values of the classifier's hyper-parameters.**

Keywords*—Machine learning; Arabic sentiment analysis; optimized multinomial Naïve Bayes (MNNB) classifier; hyper-parameters optimization

I. Introduction

Recently, sentiment analysis (SA) using machine learning methods has become an important research task for developing various text-based applications [1, 2]. It is used in natural language processing (NLP), text search, and computational linguistics for extracting people's opinions or feelings about products, events, or other in one way or another [3]. The SA imposes the identification of several elements which are four elements including the entity, its aspect, the opinion holder, and its feeling and can categorize the opinions extracted into either a subjective or objective text. The subjective text can also be categorized into negative or positive feelings. Several methods have been conducted on sentiment analysis of several languages, including Arabic, and differences have been discovered, as NLP in Arabic is still in its early steps and lacks tools and resources [4]. Hence, the Arabic language still faces difficulties in NLP tasks because of its complexity, structure, and the different dialects to which it belongs. A large number of tools and methods have been used, in the literature, to perform the sentiment analysis task. Most of them are considered in English, which is the science language, these tools and methods are based on either a machine learning approach or a semantic approach [5]. The semantic approach extracts emotion words and computes their poles based on the emotion dictionary. Conversely, in a machine learning approach to building a new model, machine learning classifiers are trained on pre-labeled data after transforming into feature vectors. Finally, the new model can be applied for predicting a new category of data based on these characteristics. It should be noted that these methods can be modified to another language, such as Arabic. The Arabic language did not receive the effort that other languages did [6]. However, several studies of sentiment analysis of Arabic writing have been proposed [7]. A decade ago, Arabic sentiment analysis became one of the most common information mining forms in many fields. These analytics have contributed to achieve many benefits, such as offering brand value insights for a product or service Invite potential product customers, identify social media influencers, and detect spam. Consequently, Arabic sentiment analysis has been investigated in different contexts, and a number of techniques in several studies have been published on this topic. However, there is still a limitation in MNNB classifier, which is widely used to analyze users' sentiment from texts and classify their topics.

This study aims to develop an optimized simple ML model-based approach for improving Arabic sentiment analysis using grid search algorithm and multinomial naïve Bayes (MNNB) classifier. The approach is able to select the optimal values for alpha hyper-parameter and control the decision boundary of the classifier. Through this proposed approach, the research contributions to the field can be presented in the following points:

- Propose an improved approach for Arabic sentiment analysis using a grid search algorithm and MNNB classifier that can be able to optimize the learning process of the classifier.
- Use the grid search algorithm as a selection step to assign the optimal or near-optimal values of MNNB classifier's alpha-parameter, improving the sentiment analysis of Arabic tweets.
- Train the proposed approach on three datasets with a large number of Arabic tweets for achieving the diversity in learning the ML models.

- Evaluate and compare the performance results of the optimized model with the baseline models on the same test dataset.

The rest of the paper is structured as follows: Section II gives a detailed literature review for the study. Section III describes the research methodology. Section IV presents the experimental results with discussions and findings. Finally, the conclusions and future work are summarized in Section V.

II. Literature Review

Nowadays, the rapid development of social media makes the text messages posted by users become the largest public data source in the world. Such text messages contain an important information and a great commercial and research value. Sentiment analysis and text analytics using NLP is one of the key methods that can provide a necessary support for text analysis in the social media. Consequently, text analytics technique including sentiment analysis, entity recognition, topic modeling, and text summarization using NLP in social media has attracted widespread attention.

The social media Arabic sentiment analysis can analyze and mine the tendency and view of user expressions from his subjective text. This analyzing supports the decision making of different researchers, users, government agencies, and business organizations. In this context, the worth question of discussion is how to effectively mine these massive textual information, identify the sentiment in it, and use it reasonably and effectively. Sentiment analysis, also known as propensity analysis, is a computational study of emotions, opinions, and feelings held by people about things and their attributes [8]. Things can be services, products, individuals, organizations, questions, events, or topics. Sentiment analysis task can also be defined as the process of automatically mining attitudes, opinions, opinions, and emotions from speech, text, Weibo, and other data sources through NLP technology [9]. Text sentiment analysis is to analyze the sentiment of a paragraph of text, as the basic work of public opinion monitoring, and has a wide range of uses.

Social networks are getting more and more popular, and "opinion leaders" are getting more and more. Sites that allow users to rate product and service evaluations have sprung up, and user reviews and suggestions can be spread throughout the network. These text-type data are undoubtedly the source of the power of precision marketing. Enterprises can build their own digital image based on sentiment analysis, identify new market opportunities, do a good job of market segmentation, and then promote the successful listing of products. But grasping the value of these reviews is also a huge challenge for companies. Governments, like enterprises, need to monitor, alleviate, and lead public opinion through sentiment analysis, and eliminate social conflicts. The above is the application background of sentiment analysis. But contrary to such an important background is the weakness of the Arabic sentiment analysis system.

Common sentiment analysis is separated into dictionary-based sentiment analysis and supervised model-based analysis. Dictionary-based sentiment-analysis, as the name suggests, relies heavily on the construction of sentiment dictionaries. Ku et al. [10] and Kaji et al. [11] conducted in-depth research on the construction of sentiment dictionaries. Generally, the emotional words are first divided into positive (meaning) and negative (derogatory), and then the number of positive words and negative words of an Arabic text to be analyzed is counted. If the positive words number is greater than the negative number of words, this text belongs to the positive emotion, otherwise it belongs to the negative emotion.

Some researchers have artificially weighted sentiment dictionaries. However, no matter how it is changed, this analysis method has the following limitations: first, the accuracy is very low, which can hardly support the requirements of public opinion monitoring; second, the positive or negative tendency or weight of emotional words is manually defined, and the workload is huge and very arbitrary; in the end, this approach is almost ineffective for negative sentences and sentences reinforced by adverbs of degree, thus losing the ability to analyze the delicateness (degree) of emotions.

Social media is an online interactive platform based on online social networks and the main form of Internet users' creation and dissemination of information. Twitter and Facebook are typical examples of online social networks. The emergence of social media has epoch-making significance, so that the general public's emotions can be easily and fully expressed, spread, and influence each other. There are a large number of research problems in the field of sentiment analysis. Many of today's naming related tasks with small differences are usually included in the research field of sentiment analysis, such as opinion mining, opinion, sentiment analysis, comment mining and so on. Text sentiment analysis [12-15] aims to analyze the attitudes and sentiment of opinion expressers, that is, to analyze the subjective information in the text.

Although text sentiment analysis studies texts with a polarity have started before the year 2000, few scholars studied the sentiment from texts in the field of linguistics and NLP. This may be partly because there was no growth in digital records at that time. With the explosive growth of the Internet and social media, people can have an uninterrupted data stream and store it in digital form, which is also an important reason why sentiment analysis has maintained a consistent growth rate with the Internet in recent years. For many years, social media systems have provided users with a very convenient channel to communicate and share. The important carrier of information is text information in social media.

People are keen to be in such a free and convenient environment that is not limited by time and space. Make your own voice, express your views on everything, and establish connections between users. This user-generated content provides researchers with great convenience to track, collect, store, retrieve, and analyze people's emotional changes. Appearance has injected new vitality into sentiment analysis, and sentiment analysis has also provided a new research area for social media analysis.

As early as 1997, Tiwari et al. [16] began to try to use the conjunctions in linguistics that are binding on sentiment words to infer the opinions and attitudes of the whole article, that is, to use adjectives with known sentiments to infer The emotional

tendency of an adjective. Turney et al. [17] used the association between words and some seed words with obvious semantic tendencies and combined statistical methods to identify the propensity characteristics of words. The research term of sentiment analysis may be first proposed in the literature of Nasukawa and Yi [14] in 2003, but many related work of sentiment analysis has been started before 2003 [15-17]. Earlier related studies include interpretation of metaphors, extraction of adjectives with emotions, sentiment calculation, subjective analysis, and so on.

The existing applications and research of sentiment analysis are mainly focused on text, which has become a hot spot in the field of NLP research. Since 2002, sentiment analysis research has become very active. In addition to the large number of trending texts in social media, its extensive application scenarios have become increasingly prominent in various human activities. This is also social media sentiment analysis is different from traditional text sentiment analysis. Social sentiment-oriented text sentiment analysis is mainly based on information sharing and interactive review mechanisms. When people need to make a decision at a certain time, most people often refer to the opinions of others. This situation is not only for individuals but also for enterprises and institutions.

Because the large amount of information with user sentiment is publicly available on the Internet, companies no longer use a large number of questionnaires to collect and understand consumer opinions on their products, and the government can easily grasp the public's perspective to supervise their regions. Therefore, social media from about 2006, sentiment analysis has ushered in its prosperity. Its widespread application has given rise to a strong demand for research, and at the same time brought many unprecedented challenges. These challenges are exactly the problems that this article needs to solve.

The related research on text sentiment analysis for social media has been widely concerned by academia and industry. Researchers and institutions have invested a lot of manpower and material resources in order to use text sentiment analysis technology to obtain relevant information. At present, text sentiment analysis has been discussed in many international top conferences.

In the industrial sector, such as major e-commerce shopping websites and portals, they have applied sentiment analysis technology to user review analysis, and found problems in the product and improved them through user reviews, thereby achieving the goal of increasing product sales.

Develop a comprehensive social media analytics tool to help decision makers with external customers to understand sentiment and feedback mapped to the services/products in discussion. The data source is social media feeds such as Twitter or Facebook. The model should be an engine to analyze Saudi dialect and English using advanced NLP algorithms which can be used with any free text platforms to understand main topics and sentiment analysis.

In recent years, Arabic sentiment analysis has become a popular research topic. Using an SVM classifier, Abdul-Mageed et al. [18] investigated the effect of subjectivity and sentiment classification at the sentence level for the Modern Standard Arabic language (MSA). Shoukry and Rafea [19] used 1000 tweets to apply SVM and Nave Bayes (NB) at the sentence level for sentiment classification. For sentiment analysis, Abdulla et al. [20] compared lexicon-based and corpus-based approaches.

The lexicon construction challenges and sentence analysis were addressed by Abdulla et al. [21]. Badaro et al. [22] used an English-based relating to a WordNet and the lexicon approach to create a large Arabic sentiment lexicon. Duwairi et al. [23] used crowdsourcing to collect over 300,000 Arabic tweets and label over 25,000 of them. For Arabic sentiment classification, Al Sallab et al. [24] applied three deep learning methods. Ibrahim et al. [25] used different types of text data, such as tweets and product reviews, to show sentiment classification for Egyptian dialect. The use of pre-trained models with CNN for Arabic word representation improved sentiment analysis performance, according to Dahou et al. [26].

Researchers developed aBidirectional LSTM Network (BiLSTM) with efficient feature extraction capability in [27]. The backward and forward dependencies are used to extract information from feature sequences. For evaluating the models' performance, a number of experiments was conducted on six benchmark datasets of sentiment analysis. The results show that the proposed model outperformed the other models, including the Support Vector Machine (SVM), Random Forest (RF), and LSTM. The authors of [28] investigated various deep learning (DL) models for sentiment analysis of Arabic microblogs based on LSTM and CNN. They used datasets from continuous bag-of-words (CBOW), skip-gram (SG), ASTD, and Ar-Twitter in their experiments. The experimental results revealed that LSTM outperforms CNN. In another work, the authors [29] used NB and SVM with different schemes for weighting such as n-gram sizes and TF-IDF to analyze the Arabic sentiment. They performed the experiments on AJGT dataset and they found the best performance is for the scenario of SVM classifier. The hybrid models-based DL algorithms for sentiment classification were proposed by the authors in [30]. They compared and evaluated the hybrid model to DT, RF, CNN, and RNN-LSTM, using over one million tweets from various domains. The hybrid model performed the best, according to the results.

To predict Arabic sentiment analysis, A. M. Alayba et al. [31] used a variety of machine learning models, including NB, SVM, Ridge Classifier (RDG), LR with Stochastic Gradient Descent (SGD), and the DL models, such as CNN and other methods of feature extraction. The Arabic Health Services Dataset was used to conduct the tests. Mohamed Fawzy et al. [32] stated a diversity of deep learning network models for classifying Arabic sentiment, as well as word embedding techniques. To conduct experiments, we used RNN, CNN, and Bidirectional Multi-Layer-LSTM with various word embedding. Large-scale Arabic book reviews were used in the experiments (LABR). The results revealed that the Bidirectional Multi-Layer LSTM has a high level of precision. In [33], the authors used NB, DT, LR, SVM, and DL models on Arabic tweets dataset for Saudi dialect sentiment analysis. Deep learning and SVM classifiers outperform all others in terms of accuracy. To predict sentiment analysis, some studies

used an ensemble learning (EL) approach. To predict sentiment analysis, Al-Hashedi et al. [34] used NB, RF, voting ensemble method, SGD, and LR. The author gathered COVID-19 Arabic tweets and classified them as negative or positive. The voting classifier performs well, as evidenced by the results. Alharbi et al. [35] proposed a DeepASA architecture model that included hidden layers, as well as input and output layers. Two types of deep neural networks (LSTM and GRU), and a voting classifier were used to improve the prediction performance of the model. Large Scale Arabic Hotel Reviews (HTL), Library Book Reviews Dataset (LABR), Product Reviews (PROD), Restaurant Reviews (RES), ASTD datasets, and ArTwitter were used in the experiments. The DeepASA-based approach performs well, as evidenced by the results. On the ASTD dataset of sentiment analysis for Arabic text, Oussous et al. [36] used a voting method built on top of three classifiers: NB, SVM, and Maximum Entropy. The results show that the vote algorithm is extremely accurate.

For classifying the Arabic text sentiment, Al-Saqqa et al. [37] offered an ensemble approach combines three ML classifiers, SVM, KNN, and NB using a majority voting algorithm. The datasets such as ArTwitter, Movie reviews, and a large-scale Arabic sentiment analysis were used in this study. The results of the experiments revealed that the ensemble of classifiers outperforms individual classifiers. On the Arabic sentiment dataset, Al-Azani et al. [38] studied the performance of various ensemble learning methods to enhance the performance of single classifiers, including boosting, bagging, stacking, voting, and RF. The stacking ensemble performs well, as evidenced by the results. Other researchers used ensemble learning techniques to analyze sentiment in other languages than Arabic language. For example, Sitaula et al. [39] created the NepCOV19Tweets, a Nepali Twitter sentiment dataset that was labeled negative, neutral, and positive. The authors developed some feature extraction methods based on fastText, domain-agnostic, and domain-specific feature selection techniques. Each feature selection method was implemented using different CNN models. Then, for capturing multi-scale information in order to obtain better classification results, they proposed a CNN ensemble model. The authors suggested a multi-channel CNNs (MCNNs) classification system for classifying the tweets in NepCOV19 dataset into negative, neutral, and positive sentiment [40]. A hybrid feature extraction method has been proposed for extracting syntactic and semantic features to train their proposed MCNNs model. When compared to single methods of feature extraction, the hybrid features achieved the highest accuracy result, and the MCNNs model obtained a best performance threshold. Ahmed Mohamed [41] applied Naive Bayes and SVM algorithms for Arabic sentiment analysis and El-Masri et al. [42] presented a novel method for sentiment analysis to Arabic language tweets that uses a mix of parameters related n-grams-based features and preprocessing methods. The tool analyzes the most recent tweets about the issue to determine the polarity (positive, negative, and neutral) and then displays the findings. The results of the study demonstrated that the Naive Bayes method is the most effective in detecting topic polarity. Expert and intermediate users can choose the most effective combinations of parameters for sentiment analysis with the aid of the tool. However, the related studies have a limitation in selecting the best values to hyper parameters of Naive Bayes method.

Among different ML classifiers used in previous studies, MNNB classifier is widely used for documents classification due to its ability for performing statistical analysis of text contents. It significantly simplifies textual-data classification and offers an alternative to heavy ML-based semantic analysis methods. However, the MNNB classifier has a number of hyper-parameters affects the classification task of texts and controls the decision boundary of the model itself. In this paper, an optimized MNNB classifier-based approach is proposed for improving Arabic sentiment analysis.

III. Research Methodology

This section describes the research methodology of proposed approach for social media Arabic tweets analytics. It is based on an optimized MNNB classifier model. The steps of the research methodology are shown in Fig. 1.

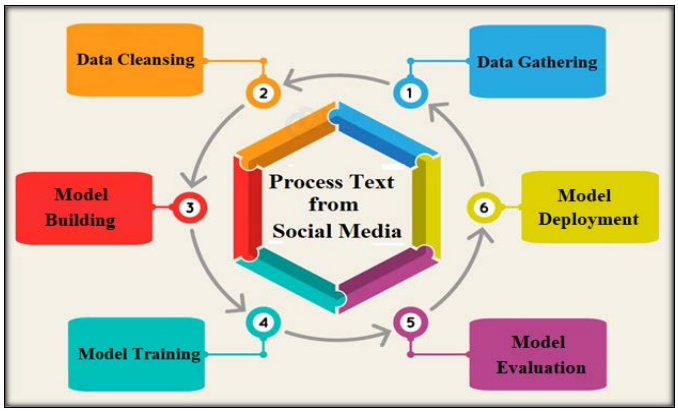


Fig. 1. Research Methodology.

A. Data Gathering

As mentioned in the scope of this research work, the model should be able to analyze Arabic sentiment from users' tweets in the highly used social networks, Twitter. The data gathered are the tweets and their retweets. Then, the data collected from the Twitter source will be stored in a dataset for training after manually labeling or for testing in the evaluation phase. Once gathering a list of Arabic topics or keywords, you can find them on Twitter and then save their related information in the storage device for further processing. The Twitter API makes it simple to search for users and returns results in JSON format. This format is easy to parse in a Python script. Dealing with social media accounts may face one complication, which is the fake accounts with similar or identical names, making them difficult to detect. Fortunately, each user object in Twitter includes a handy data field that indicates whether the account is verified, which I checked before saving the handle. The next step was to use Twitter's API to download the user's tweets and save them into a dataset once a topic or required keyword was linked to a Twitter handle.

B. Data Cleansing

It is a part of NLP since the data received contains incomplete, incorrect, inaccurate or irrelevant parts of the data records that required replacing, modifying, or deleting them.

Therefore, it will be used in the proposed methodology for English and Arabic text. It can perform the following:

- Delete the empty records.
- Delete the retweets.
- Remove the Hashtags, pictures, and links.
- Remove usernames from mentions.
- Remove English and Arabic stop words.
- Normalize the words.
- Stemming the words.

After performing the previous functions, the data will be ready to be used in the model for training and testing purposes.

C. Model Building

After data cleansing, the selection of a model is the next step in the research methodology process. Over the years, researchers and data scientists have developed a variety of models. Some are better suited to image data, while others are better suited to sequences (such as voice or text), numerical data, and text-based data. A multinomial naïve Bayes (MNNB) classifier is a simple and effective ML model to deal with text features. The MNNB classifier is often used as a starting point for sentiment analysis.

The core idea behind the MNNB technique is to use the joint probabilities of words and classes to find the probabilities of classes given to texts.

Given a vector of dependent features $(f_1, \dots, f_n)$ and the class L_k, Bayes' theorem can be expressed mathematically as:

$$P(L_k|f_1, \dots, f_n) = \frac{P(L_k)P(f_1,\dots,f_n|L_k)}{P(f_1,\dots,f_n)} \quad (1)$$

For the given class L_k and consistent with the assumptions of naïve conditional independence, each feature f_i is conditionally independent of every other feature f_j where $i \neq j$.

$$P(f_i|L_k, f_1, \dots, f_n) = P(f_i|L_k) \quad (2)$$

Thus, it can be simplified the relation to be as:

$$P(L_k|f_1, \dots, f_n) = \frac{P(L_k)\prod_{i=1}^{n} P(f_i|L_k)}{P(f_1,\dots,f_n)} \quad (3)$$

Because $P(f_1, \dots, f_n)$ is constant and if the feature values of the known variables, the following rule for classification can be employed:

$$P(L_k|f_1, \dots, f_n) \propto P(L_k)\prod_{i=1}^{n} P(f_i|L_k)$$

$$\Downarrow$$

$$\hat{y} = arg\max_k P(L_k)\prod_{i=1}^{n} P(f_i|L_k) \quad (4)$$

Log probabilities can be used to avoid underflow.

$$\hat{y} = arg\max_k(\ln P(L_k) + \sum_{i=1}^{n} \ln P(f_i|L_k)) \quad (5)$$

According to the distribution of $P(f_i|L_k)$, the assumptions made by the Naive Bayes classifier differs between them, whereas $P(L_k)$ is relatively defined as the frequency of class L_k in the training data set.

The distribution of multinomial naïve Bayes is parametrized by the vector $\theta_k = (\theta_{k1}, \dots, \theta_{kn})$ for each class L_k , where n represents the number of features (vocabulary size) and θ_{ki} denotes the probability $P(f_i|L_k)$ of feature i that appears in a sample that belongs to the class L_k.

The estimation of parameters θ_k can be obtained by a smoothed version of maximum likelihood (i.e., relative frequency counting) as follows:

$$\hat{\theta}_{ki} = \frac{N_{ki}+\alpha}{N_k+\alpha n} \quad (6)$$

Where N_{ki} represents the number of times feature i appears in class a sample k of the training dataset T. The total count of all features for class L_k is N. The smoothing parameter alpha (α) can have a value greater than zero and less than or equal 1 $0 < \alpha \leq 1$ for features, which are not present in the learning samples and to prevent from division by zero probabilities in further computations.

Setting $\alpha = 1$ is termed as Laplace smoothing, while $\alpha < 1$ is termed as Lidstone smoothing. Therefore, the final decision rule is written as:

$$\hat{y} = arg\max_k(\ln P(L_k) + \sum_{i=1}^{n} \ln \frac{N_{ki}+\alpha}{N_k+\alpha n}) \quad (7)$$

Unfortunately, the way to know what α gives the most accurate responses is through iterating over all values of α on the training set and this way is a hard problem. The optimized MNNB classifier used in this research determines the optimal value for hyper-parameters is through a grid search over possible parameter values and using cross-validation evaluation technique for each value. The flowchart of the optimization process for MNNB classifier-based proposed approach is shown in Fig. 2.

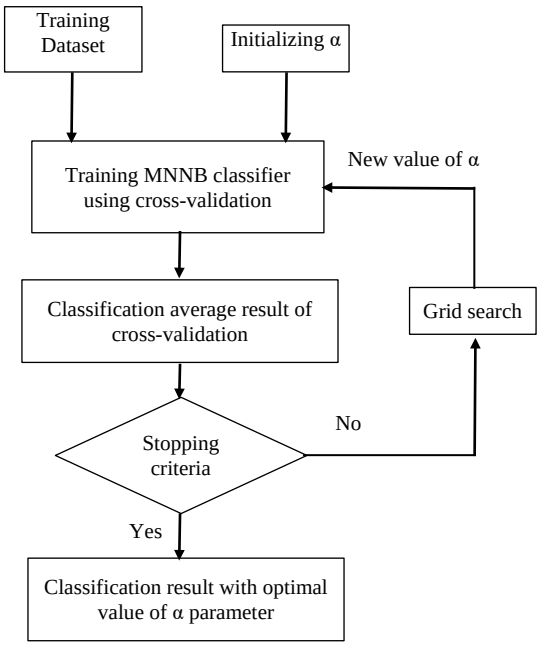


Fig. 2. Flowchart of Optimized MNNB Classifier-based Approach.

D. Model Training

The training phase is where the main process of developing the baseline and proposed models begins. Three different datasets are used for training the models of the research approach. By analyzing the characteristics of Arabic tweets, the models learn to classify their sentiments. For the MNNB classifier, the model is trained through mining the statistics from the training dataset. The utilization of frequency and likelihood tables for each feature facilitates the calculation process.

Possible feature values are grouped together in a frequency table derived from the observations. Instead of counting the number of occurrences, the likelihood table shows the probability values for each class. The multiplication and comparison operations are used to determine the class of an observation using these tables. The training process of MNNB classifier can be summarized in the following steps:

- Convert the Arabic text data into words vectors such as using TFIDF vectorizer or count vectorizer.
- Calculate the counts based on the class.
- Calculate all the likelihood probabilities.
- Calculate the prior probability.
- Calculate the posterior probability.

Feature vectors represent the frequency with which specific events were generated by a multinomial distribution. This is the most commonly used event model for Arabic text classification. This algorithm is used to solve problems with Arabic text classification. This method could be used to determine whether a tweet belongs in the ‘positive’ or ‘negative’ category, for example. It takes advantage of the current words’ frequency as a feature.

E. Model Evaluation

The model must be tested after it has been trained with train data. The goal of testing is to see how the model performs in real-world situations. During this phase, we can assess the model's accuracy. In our case, the model uses the learning from the previous phase to try to identify the type of fruit. The evaluation phase is crucial, as it allows us to see if the model achieves the goal we set for it. If the model does not perform as expected during the testing phase, the previous steps must be repeated until the required accuracy is achieved. As stated, it should not use the same data that was used in the training phase. For evaluation, it should have to use the separate data splitter from the dataset.

The only thing that classification models care about is whether or not the result is correct. When making classification predictions, such as the one we used, there are four possible outcomes. True negatives, true positives, false negatives, and false positives are the four types. On a confusion matrix, these four outcomes are plotted. After making predictions on the test data, you can create the matrix and categorize each prediction as one of the possible outcomes. The percentage of correct predictions made by the test data determines the model's accuracy. The model's accuracy can be determined by dividing the number of correct predictions by the total number of predictions. Classification models are also evaluated using other metrics such as accuracy.

F. Model Deployment

Model deployment is the integrating process of ML model in an existing construction environment to make decisions for data-driven business. It is the last step in the ML process and one of the most time consuming steps. Traditional model-building languages are frequently incompatible with an organization's IT systems, which force programmers and data scientists to spend brainpower and valuable time for rewriting them.

A model must be successfully deployed into production before it can be used for practical decision-making. If you can't rely on your model to provide practical insights, then its impact is severely limited. One of the difficult aspects of achieving value from ML is model deployment. IT teams, software developers, data scientists, and business professionals must work together to ensure that the model works reliably in the organization's production environment. This is a significant challenge due to there is frequently a mismatch between the programming-language used to create an ML model and the languages that the production system understands. Model re-coding can add weeks or months to the project timeline.

The deployment of ML models is the final step in the process. ML models are typically developed and tested using training and testing datasets in a local or offline environment. When a model is deployed, it is placed in a live environment and is exposed to new and unknown data. As the model performs the task, it was trained for working on live data, the model begins to bring to the organization a return on investment.

Containerization is becoming increasingly popular as a tool for deploying ML models. Containers are a common environment to deploy the models because they simplify updating and deploying different parts of the model. Containers are intrinsically scalable, as well as able to provide a consistent environment for the model function. Kubernetes and other open-source platforms are utilized for managing and automating the container management aspects such as scaling and scheduling.

IV. Experimental Results and Discussions

Before presenting the experimental results, this section starts by presenting the number of instances for each class in the training and test datasets. The adopted ML models are trained on three different datasets. Each dataset has a larger number of tweets as shown in Fig. 3 to 5. For the dataset 1, Fig. 3 illustrates that the number of negative instances is 976, and the positive instances is 1046, as well as the number of natural instances is 724. For the dataset 2, Fig. 4 demonstrates that the number of negative instances is 2588, and the positive instances are 1817, and the number of natural instances is 1587. Similarly, the dataset 3 contains 228 negative instances, 263 positive instances, and 192 natural instances as seen in Fig. 5. The test dataset is also used to evaluate the trained models for sentiment classification of Arabic tweets. Fig. 6 displays the number of instances in the test dataset.

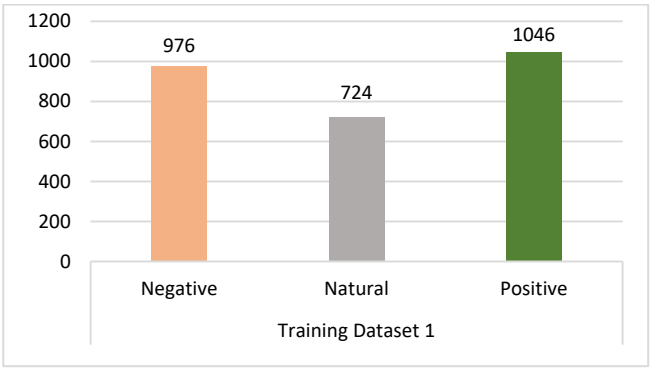


Fig. 3. Number of Instances in Training Dataset 1.

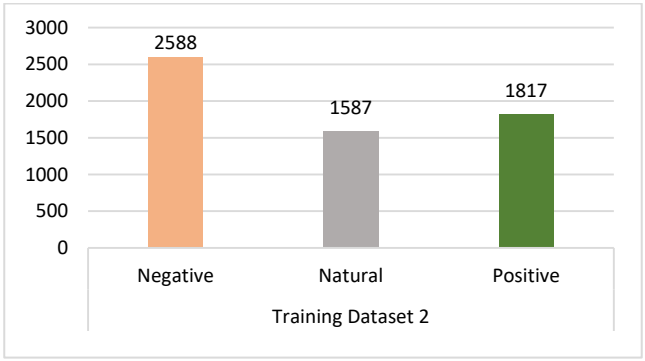


Fig. 4. Number of Instances in Training Dataset 2.

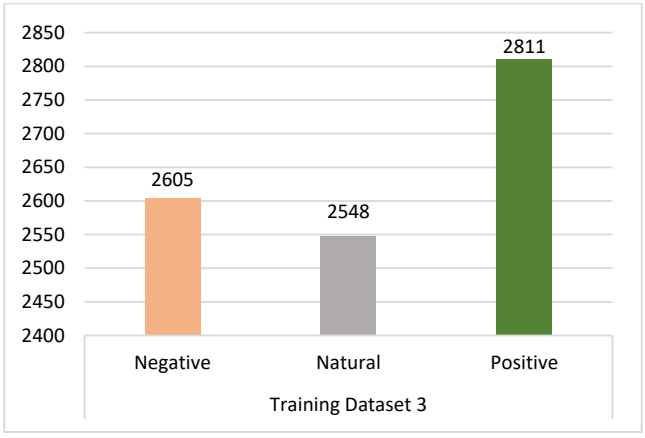


Fig. 5. Number of Instances in Training Dataset 3.

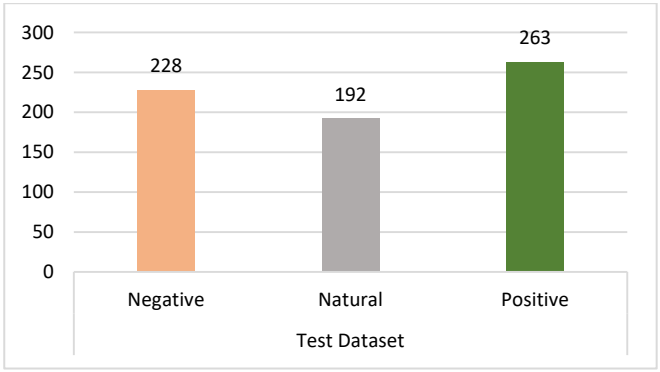


Fig. 6. Number of Instances in Test Dataset.

From Table I to Table VI, the confusion matrices of test set classification for the optimized MNNB trained on the three datasets is given using TFIDF and count vectorizers. Table VII and Table VIII show the evaluation results, which show each classifier's performance on positive, natural and negative instances, as well as their overall performance.

TABLE I. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 1 AND USING TFIDF VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	139	15	74
Natural	71	37	84
Positive	44	19	200

TABLE II. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 1 AND USING COUNT VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	136	25	67
Natural	62	57	73
Positive	49	23	191

TABLE III. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 2 AND USING TFIDF VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	167	12	49
Natural	109	28	55
Positive	73	19	171

TABLE IV. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 2 AND USING COUNT VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	136	26	66
Natural	77	50	65
Positive	53	30	180

TABLE V. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 3 AND USING TFIDF VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	111	39	78
Natural	72	46	74
Positive	33	36	194

TABLE VI. CONFUSION MATRIX OF TEST DATASET FOR THE OPTIMIZED MNNB CLASSIFIER TRAINED ON DATASET 3 AND USING COUNT VECTORIZER

Actual / Predicted	Negative	Natural	Positive
Negative	115	42	71
Natural	69	54	69
Positive	34	36	193

TABLE VII. CLASSIFICATION ACCURACY RESULT USING TFIDF VECTORIZER

Classifier	Training Dataset 1	Training Dataset 2	Training Dataset 3
SVM	0.511	0.511	0.515
SVM Linear Kernel	0.530	0.531	0.518
RF	0.488	0.486	0.463
GaussianNB	0.483	0.474	0.466
MNNB	0.539	0.521	0.514
Optimized MNNB	**0.551**	**0.536**	**0.514**

TABLE VIII. CLASSIFICATION ACCURACY RESULT USING COUNT VECTORIZER

Classifier	Training Dataset 1	Training Dataset 2	Training Dataset 3
SVM	0.492	0.476	0.460
SVM Linear Kernel	0.526	0.518	0.508
RF	0.488	0.502	0.466
GaussianNB	0.482	0.480	0.473
MNNB	0.546	0.542	0.530
Optimized MNNB	**0.562**	**0.542**	**0.530**

From the above results, the following observations are made:

- From the three datasets used for training as shown in Fig. 3 to 5, we can see that there is a diversity in the number of instances for each class to see the effect of class size for training the models.
- From Table I to Table VI, the confusion matrices show that the number of instances, which are correctly classified for the optimized MNNB classifier is improved by selecting the optimal values of α hyper-parameter, especially on neutral and negative instances.
- As shown in Table VII and Table VIII, the optimized MNNB classifier obtains high accuracy results by using count vectorizer representation for Arabic tweets as features rather than TFIDF vectorizer.
- SVM with linear kernel works well than other models, especially using TFIDF vectorizer but not better than MNNB and optimized MNNB classifiers.
- For sentiment classification of Arabic tweets, MNNB model is preferable as a generative model. It outperforms the other baseline classifiers.

V. CONCLUSION AND FUTURE WORK

Arabic sentiment analysis using machine learning methods has become an important research task for developing various applications. In this paper, an optimized MNNB classifier-based approach is presented for improving Arabic sentiment analysis. It aims to select the optimal value of the MNNB's alpha hyper-parameter to control the decision boundary of the model itself. The sentiment classification experiments of the research are conducted using a large-scale data sets. The baseline and optimized MNNB classifiers are trained on three datasets and tested on a different separated test set to show the performance of developed approach. The experimental results on the test set revealed that the optimized MNNB classifier of proposed approach outperforms the traditional MNNB classifier and other baseline classifiers. The accuracy rate of the optimization approach is increased by 1.6% compared with using the default values of the classifier's hyper-parameters. The output from the study shows that a MNNB classifier with count vectorizer as features can achieve a high performance compared to the other baseline classifiers. Because there are a large number of Arabic tweets features that are likely to be noisy, a feature selection scheme can be investigated in future work. Previous classification studies have shown that feature selection is critical for classification task success. Another promising extension of this research would be to classify Arabic tweets on different scales instead of just positive, neutral, and negative classes. Moreover, a combination of different classifiers and deep learning methods will be explored.

ACKNOWLEDGMENT

"The author is thankful to the Deanship of Scientific Research, College of Computer and Information Sciences (CCIS) at King Saud University for funding this research."

REFERENCES

[1] M. Birjali, M. Kasri and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," Knowledge-Based Systems, vol. 226, p. 107134, 2021.

[2] D. Antonakaki, P. Fragopoulou and S. Ioannidis, "A survey of twitter research: Data model, graph structure, sentiment analysis and attacks," Expert Systems with Applications, vol. 164, p. 114006, 2021.

[3] S. Zad, M. Heidari, J. H. Jones and O. Uzuner, "A survey on concept-level sentiment analysis techniques of textual data," in 2021 IEEE World AI IoT Congress (AIIoT), 2021, pp. 0285-0291: IEEE.

[4] M. O. Hegazi, Y. Al-Dossari, A. Al-Yahy, A. Al-Sumari and A. Hilal, "Preprocessing arabic text on social media," Heliyon, vol. 7, no. 2, p. e06191, 2021.

[5] R. Bensoltane and T. Zaki, "Aspect-based sentiment analysis: An overview in the use of arabic language," Artificial Intelligence Review, pp. 1-39, 2022.

[6] M. Hijjawi and Y. Elsheikh, "Arabic language challenges in text based conversational agents compared to the english language," International Journal of Computer Science Information Technology, vol. 7, no. 5, pp. 1-13, 2015.

[7] S. L. Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali and I. Abunadi, "Arabic natural language processing and machine learning-based systems," IEEE Access, vol. 7, pp. 7011-7020, 2018.

[8] B. Liu, "Sentiment analysis and subjectivity," Handbook of natural language processing, vol. 2, no. 2010, pp. 627-666, 2010.

[9] V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: A survey of techniques," arXiv preprint arXiv:.06971, 2016.

[10] L.-w. Ku, Y.-s. Lo and H.-h. Chen, "Using polarity scores of words for sentence-level opinion extraction," in Proceedings of NTCIR-6 workshop meeting, 2007, pp. 316-322: Citeseer.

[11] N. Kaji and M. Kitsuregawa, "Building lexicon for sentiment analysis from massive collection of html documents," in Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007, pp. 1075-1083.

[12] L. Yang, B. Liu, H. Lin and Y. Lin, "Combining local and global information for product feature extraction in opinion documents," Information Processing Letters, vol. 116, no. 10, pp. 623-627, 2016.

[13] L. Yang, H. Lin, Y. Lin and S. Liu, "Detection and extraction of hot topics on chinese microblogs," Cognitive Computation, vol. 8, no. 4, pp. 577-586, 2016.

[14] W. Han, Z. Tian, Z. Huang, S. Li and Y. Jia, "Topic representation model based on microblogging behavior analysis," World Wide Web, vol. 23, no. 6, pp. 3083-3097, 2020.

[15] L. Huang, S. Li and G. Zhou, "Emotion corpus construction on microblog text," in Workshop on Chinese Lexical Semantics, 2015, pp. 204-212: Springer.

[16] P. Tiwari, P. Yadav, S. Kumar, B. K. Mishra, G. N. Nguyen et al., "Sentiment analysis for airlines services based on twitter dataset," Social Network Analytics: Computational Research Methods & Techniques, vol. 149, 2018.

[17] L. Yang, S. Zhang, H. Lin and X. Wei, "Incorporating sample filtering into subject-based ensemble model for cross-domain sentiment classification," in Chinese computational linguistics and natural language processing based on naturally annotated big data: Springer, 2015, pp. 116-127.

[18] M. Abdul-Mageed, M. Diab and M. Korayem, "Subjectivity and sentiment analysis of modern standard arabic," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 587-591.

[19] A. Shoukry and A. Rafea, "Sentence-level arabic sentiment analysis," in 2012 international conference on collaboration technologies and systems (CTS), 2012, pp. 546-550: IEEE.

[20] N. A. Abdulla, N. A. Ahmed, M. A. Shehab and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in 2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT), 2013, pp. 1-6: IEEE.

[21] N. Abdulla, S. Mohammed, M. Al-Ayyoub and M. Al-Kabi, "Automatic lexicon construction for arabic sentiment analysis," in 2014 International Conference on Future Internet of Things and Cloud, 2014, pp. 547-552: IEEE.

[22] G. Badaro, R. Baly, H. Hajj, N. Habash and W. El-Hajj, "A large scale arabic sentiment lexicon for arabic opinion mining," in Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP), 2014, pp. 165-173.

[23] R. M. Duwairi, R. Marji, N. Sha'ban and S. Rushaidat, "Sentiment analysis in arabic tweets," in 2014 5th international conference on information and communication systems (ICICS), 2014, pp. 1-6: IEEE.

[24] A. Al Sallab, H. Hajj, G. Badaro, R. Baly, W. El-Hajj et al., "Deep learning models for sentiment analysis in arabic," in Proceedings of the second workshop on Arabic natural language processing, 2015, pp. 9-17.

[25] H. S. Ibrahim, S. M. Abdou and M. Gheith, "Sentiment analysis for modern standard arabic and colloquial," arXiv preprint arXiv:.03105, 2015.

[26] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud and P. Duan, "Word embeddings and convolutional neural network for arabic sentiment classification," in Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers, 2016, pp. 2418-2427.

[27] H. Elfaik, "Deep bidirectional lstm network learning-based sentiment analysis for arabic text," Journal of Intelligent Systems, vol. 30, no. 1, pp. 395-412, 2021.

[28] S. Al-Azani and E.-S. M. El-Alfy, "Hybrid deep learning for sentiment polarity determination of arabic microblogs," in International Conference on Neural Information Processing, 2017, pp. 491-500: Springer.

[29] K. M. Alomari, H. M. ElSherif and K. Shaalan, "Arabic tweets sentimental analysis using machine learning," in International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2017, pp. 602-610: Springer.

[30] M. H. Abd El-Jawad, R. Hodhod and Y. M. Omar, "Sentiment analysis of social media networks using machine learning," in 2018 14th international computer engineering conference (ICENCO), 2018, pp. 174-176: IEEE.

[31] A. M. Alayba, V. Palade, M. England and R. Iqbal, "Improving sentiment analysis in arabic using word representation," in 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), 2018, pp. 13-18: IEEE.

[32] M. Fawzy, M. W. Fakhr and M. A. Rizka, "Word embeddings and neural network architectures for arabic sentiment analysis," in 2020 16th International Computer Engineering Conference (ICENCO), 2020, pp. 92-96: IEEE.

[33] M. E. M. Abo, N. Idris, R. Mahmud, A. Qazi, I. A. T. Hashem et al., "A multi-criteria approach for arabic dialect sentiment analysis for online reviews: Exploiting optimal machine learning algorithm selection," Sustainability, vol. 13, no. 18, p. 10018, 2021.

[34] A. Al-Hashedi, B. Al-Fuhaidi, A. M. Mohsen, Y. Ali, H. A. Gamal Al-Kaf et al., "Ensemble classifiers for arabic sentiment analysis of social network (twitter data) towards covid-19-related conspiracy theories," Applied Computational Intelligence Soft Computing, vol. 2022, 2022.

[35] A. Alharbi, M. Kalkatawi and M. Taileb, "Arabic sentiment analysis using deep learning and ensemble methods," Arabian Journal for Science Engineering, vol. 46, no. 9, pp. 8913-8923, 2021.

[36] A. Oussous, A. A. Lahcen and S. Belfkih, "Impact of text pre-processing and ensemble learning on arabic sentiment analysis," in Proceedings of the 2nd International conference on networking, information systems & security, 2019, pp. 1-9.

[37] S. Al-Saqqa, N. Obeid and A. Awajan, "Sentiment analysis for arabic text using ensemble learning," in 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), 2018, pp. 1-7: IEEE.

[38] S. Al-Azani and E.-S. M. El-Alfy, "Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text," Procedia Computer Science, vol. 109, pp. 359-366, 2017.

[39] C. Sitaula, A. Basnet, A. Mainali and T. B. Shahi, "Deep learning-based methods for sentiment analysis on nepali covid-19-related tweets," Computational Intelligence and Neuroscience, vol. 2021, 2021.

[40] C. Sitaula and T. B. Shahi, "Multi-channel cnn to classify nepali covid-19 related tweets using hybrid features," arXiv preprint arXiv:.10286, 2022.

[41] A. Mohamed, "Svm and naive bayes for sentiment analysis in arabic," PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-1631367/v1], 2022.

[42] M. El-Masri, N. Altrabsheh, H. Mansour and A. Ramsay, "A web-based tool for arabic sentiment analysis," Procedia Computer Science, vol. 117, pp. 38-45, 2017.

AUTHORS' PROFILE

Dr. Ahmed Alsanad, is an Associate Professor of Information System Department and chair member of Pervasive and Mobile Computing, CCIS, at the King Saud University, Riyadh, KSA. He received his Ph.D. degree in Computer Science from De Montfort University, Unit Kingdom in 2013. His research interests include Cloud Computing, Health Informatics, ERP and CRM. He has authored and co-authored more than 12 publications including refereed IEEE/ACM/Springer journals, conference papers, and book chapters.

Erratic Navigation in Lecture Videos using Hybrid Text based Index Point Generation

Geeta S Hukkeri[1], R. H. Goudar[2]
Research Scholar[1], Associate Professor[2]
Department of CSE, VTU
Belagavi, India

***Abstract*—The difficulty in lecture videos is an erratic navigation in lecture video for watching only the needed portion of video content. Machine learning technologies like Optical Character Recognition and Automatic Speech Recognition allows to easily fetch the information that is hybrid text from lecture slides and audio respectively. This paper presents three main analysis for hybrid text retrieval, which is further useful for indexing the video. The experimental results indicate that the key frame extraction accuracy is 94 percent. The accuracy of the Slide-To-Text conversion achieved by this study's evaluation of the text extraction capability of Tesseract, Abbyy Finereader, Transym, and the Google Cloud Vision Optical Character Recognition is 92.0%, 90.5%, 80.8%, and 96.7% respectively. Similarly the result of title identification is about 96 percent. To extract the speech text three different APIs are used namely, Microsoft, IBM, and Google Speech-to-Text API. The performance of the transcript generator is measured using Word Error Rate, Word Recognition Rate, and Sentence Error Rate metrics. This paper found that Google Cloud Vision Optical Character Recognition and Google Speech-to-Text API have achieved best results compared to other methods. The results obtained are very good and agreeable, therefore the proposed methods can be used for automating the lecture video indexing.**

Keywords*—Automatic speech recognition; indexing; key-frames; lecture video; optical character recognition; title identification; text extraction

I. INTRODUCTION

The learning style of each individual learner has changed due to the lot of improvement in lecture videos as distance learning gives flexibility to access it independent of learner's time and place. Though the lecture recordings are convenient to learn from any place at any time there is a problem of watching only the needed topic from the long lecture recording. The focus of this paper is to generate the index points for non-linear navigation based on hybrid text. The processing of hybrid text extraction includes three analysis like "Visual screen analysis, Video OCR analysis, and Speech–to–text (STT) analysis."

Text in video pictures can be utilized as an indexing reason. Thus it is generally fair to initially identify elements from pictures. At the point when items have been effectively separated from their experiences, they likewise should be explicitly recognized. In this paper, a technique is presented that at the same time names contour and elements in binary images. There are numerous strategies that utilize certain contour highlights for ranking characters. The presented strategy marks every element utilizing a contour tracing method. A frame differencing method is used to obtain the key-frames [2]. Once the key-frames are retrieved, an OCR technique is used to extract the text from it. Current video OCR (Optical Character Recognition) methods depend on the mix of complex pre-handling methods for text extraction and conventional OCR engines. For video OCR, first video outlines must be recognized that acquire noticeable printed data; at that point, the content must be confined, the meddling background must be taken out, and mathematical changes must be applied before standard OCR engines can handle the content effectively and it is very powerful [3]. The general video OCR system comprises two fundamental advances: text detection and text recognition. Text detection measure decides the area of text inside the video picture. Microsoft Cognitive Services and Google Vision API [5] are some minimal expense answers that are presently available. The present status of technology says that recognition of object can be done using Convolutional Networks or Selective Search [7] Likewise, recognition of face is done using Fisherfaces or EigenFaces [8]. Google takes these methods and implement its AI cycles to further develop them. Google's Cloud Vision (GCV) is based on incredible PC vision models that power various Google administrations. Thus, a GCV OCR is applied to obtain sequence of strings from the key-frames.

Automatic Speech Recognition (ASR) [24] is being used in day-to-day applications. "The goal of speech recognition is to enable the humans and computers to have natural communication via speech". The accuracy of the model performance can be known with the results of transcription and segmentation obtained by the manual and automatic methods. The limitations of manual transcription such as costly, delayed performance, and error-prone when thousands of speech files are involved, lead to adaptation of automatic transcription; thus the study suggests to go for automated approach. The systems like automatic speech recognition (ASR) and text-to-speech (TTS) are performing in excess of 90% of accuracies. With AI ASR systems can result high-quality transcripts and with the usage of multi-modal data accuracy can be improved. Other than speech clarity there are many causes for the result of ASR system error rate [9] [10]. Now Google provides improved speech recognition with the usage of new technologies like "Voice Search on mobile, Voice Input, Goog411, Voice Actions, Voice Search on desktop, etc." The following are the objectives of this study:

- Retrieval of key-frames by analysing the visual screen.

- Video OCR analysis to identify title lines and extract text from lecture slides which is further helpful for creating index points.
- To extract the audio portion from the lecture videos to convert the instructor's spoken remarks into text for the purpose of creating index points.
- OCR and ASR performance is compared in order to determine which method is more effective.

II. Related Work

Effective searching and navigation of lecture video topics was especially intended for Slide Based Lecture Videos (SBLV) [11] that addresses a critical part of online talk recordings. A design for a successful Video Summarization [12] similarly as video motion rundown was proposed. A procedure for key edge removal was intended [13, 14] ward on the square based Histogram qualification and edge matching rate. Static video synopsis is perceived as a compelling style for users to rapidly peruse and understand enormous quantities of recordings [15]. Hence static video outline is considered as a clustering issue. Video skimming [16] normally viewed as a significant system for video summary. Generally, video text is installed in an exceptionally heterogeneous foundation with an incredible assortment of differences, which makes it hard to be perceived by standard OCR programming. GCV (Google cloud Vision) OCR is one of the popular API used in this case. GCV API enables the improvements of appliances that require AI help, notably for pictures comprehension [17, 18, 19]. A few examinations have successfully been made utilizing Cloud Vision API. Paper [20] efficiently executed GCV API for content-Based Image Retrieval. Mulfari carried out OCR capacity of GCV API to invent helpful innovation for humans with lack of ability, particularly for people who are evidently disabled or visually impaired. They removed text from pictures at that point express it through installed text-to-speech programming. From the study of previous works on GCV API we found that GCV API has been shown to offer one of trustworthy OCR appliance. Hence, its skill will also be examined in separating text from the talk video pictures.

For large vocabulary speech recognition, a DBN (Deep Boltzmann Machine) with a pre-trained ANN/HMM (Artificial Neural Network/Hidden Markov model) can be used. For recognizing disordered speech of the user a VIVOCA (Voice-Input Voice-Output Communication Aid) was evaluated [21], using this users can produce understandable speech from disordered speech. An Android-based application was developed for English learning using the Google Speech API, which has motivated authors to work on a speech recognition application in order to get text from the audio. The developers of the speech recognition system are expected to select Open API for the development of application speech recognition system [22]. As the study recommended to use automatic speech recognition, there is a demand for the efficiency and accuracy. Among most eminent Automatic Speech Recognition (ASR) systems, three are benchmarked on the bases of their performance namely the Google, wit, and IBM Watson [23], among these three the results of Google's ASR is better [4] [23]. The comparative study between Google Speech with Pocketsphinx shows that the background noise filtering result of Google Speech is more impressive than Pocketsphinx. According to research [1],[6], we note that the Google's Speech-to-Text outperformed other services and it has the less error rate in any case.

III. Hybrid Text Extraction from Lecture Video Images

The framework for hybrid text extraction has been shown in the Fig. 1, which includes "Visual screen analysis, Video OCR analysis, and Speech–to–text (STT) analysis." The two main parts in Lecture Video (LV) are slides (visual content) and audio (explanation of slides given by the instructor).

With the visual screen analysis we have segmented input lecture video to extract the slides from video and applied frame-differencing method to obtained key-frames. With the Video OCR analysis we have performed text detection and recognition using OCR to extract the texts from the slides. Then identified the title line from the text bounded images using geometrical information. With Speech–to–text (STT) analysis, extracting text from the audio track of the LV using ASR technology. Text from speech is one of the principle wellsprings of data in a talk video. The teacher gives detailed data about the point in the video address. The speech text is generous and unconstrained. The speech is one of the significant variables in content-based recovery of a point in a long video address.

A. Visual Screen Analysis

The main aim of this analysis is to extract key-frames from the LV. Extraction of key-frames for different sorts of video should be possible by consolidating distinctive sort of video content. Keypoint-based framework was intended to address the keyframe assurance issue with the objective that close by features can be used in picking keyframes. Usually, picked keyframes ought to be both descriptive of video content and containing least abundance. At first the long lecture video is divided into number of segments. Normally, a video holds 24 frames in a second, among them, most of the frames are repetitive. Thus, it is necessary to extract only useful frames. Generally, a video of M minutes is divided into:

$$F = M * 60 * 30 \text{frames} \quad (1)$$

where F is variety of frames created from the video at the start.

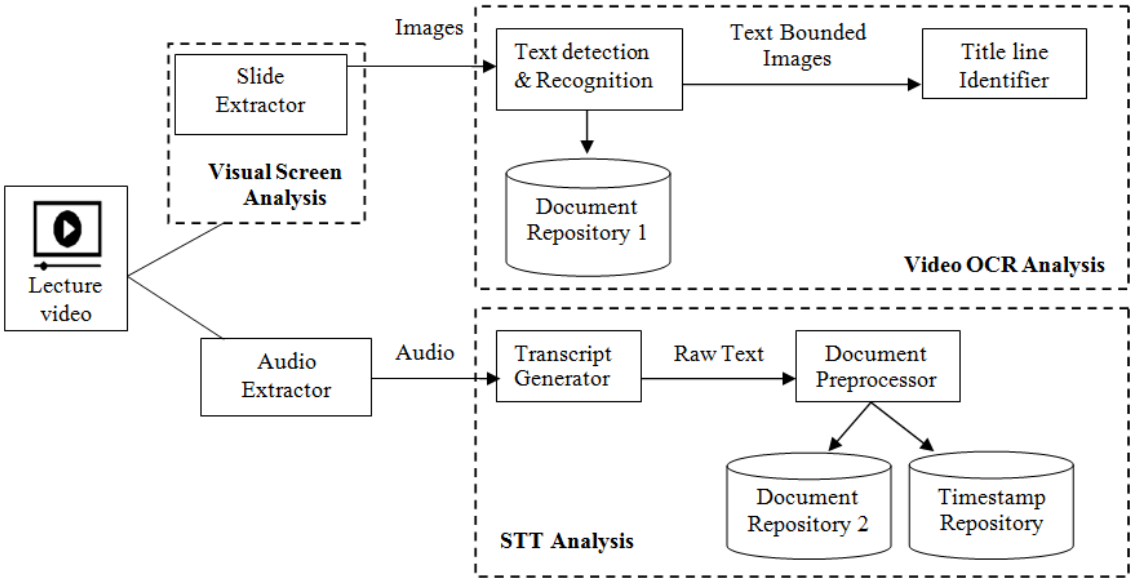


Fig. 1. Framework for Hybrid Text Extraction from Lecture Video.

In a lecture video, a topic will be deliberated for at least 10 seconds. Thus, to reduce the repetitive frames, 10 seconds delay is made in frame creation. The difference between adjacent frames is obtained to get the key-frames. The flow diagram of this procedure is shown in the Fig. 2.

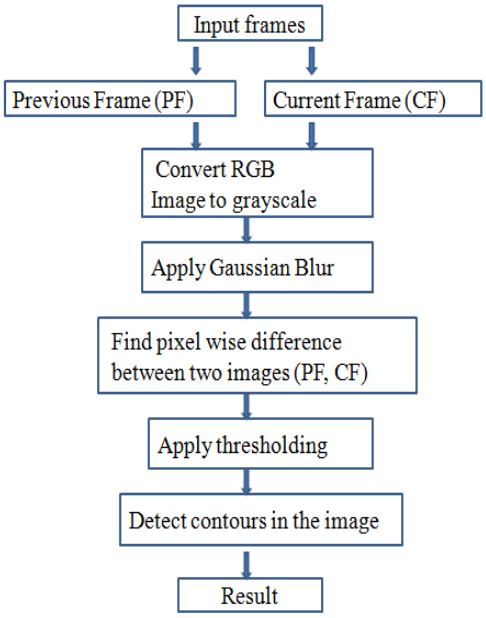


Fig. 2. Flow of Key-Frame Selection.

The method we developed for obtaining key-frames works in three stages. Initial stage is to get the difference between adjacent frames using frame-differencing method.

To compute this, frames are removed at 1 Hz from the informant video. Then compute the pixels whose difference beats the threshold value of 24. The bounding box is found for all such pixels whose change is higher than 1% of the total and analyse its size and overlay at the center of frame. A new segment is detected when the bounding box is moreover overlapping the center or is at least a third of the frame. After a time when the inter-frame difference gets steady for somewhat three seconds, a fresh key-frame is obtained as the final frame in the segment. Next, we test frames to build training sets for representing the talker/background, and slide images. Histograms are removed from the tested frames and train a SVM to discriminate slide and non-slide frames. Sample result of key-frames obtained from lecture video has been shown in the Fig. 3.

B. *Video OCR Analysis*

This section discusses two tasks.

1) Text detection and recognition: Text extraction process includes two subprocesses namely text detection and recognition which can be performed automatically by applying GCV OCR. GCV OCR is a part of GCV API. The GCV API permits developers to know the subject of a picture by enclosing wonderful AI designs in a user-friendly REST API. The Cloud Vision API rapidly groups pictures into many classes and peruses printed words contained inside pictures; it also recognizes singular items inside pictures. The Google permits the API to handle singular bits of a picture independently and return the outcome rapidly in brought together organizations.

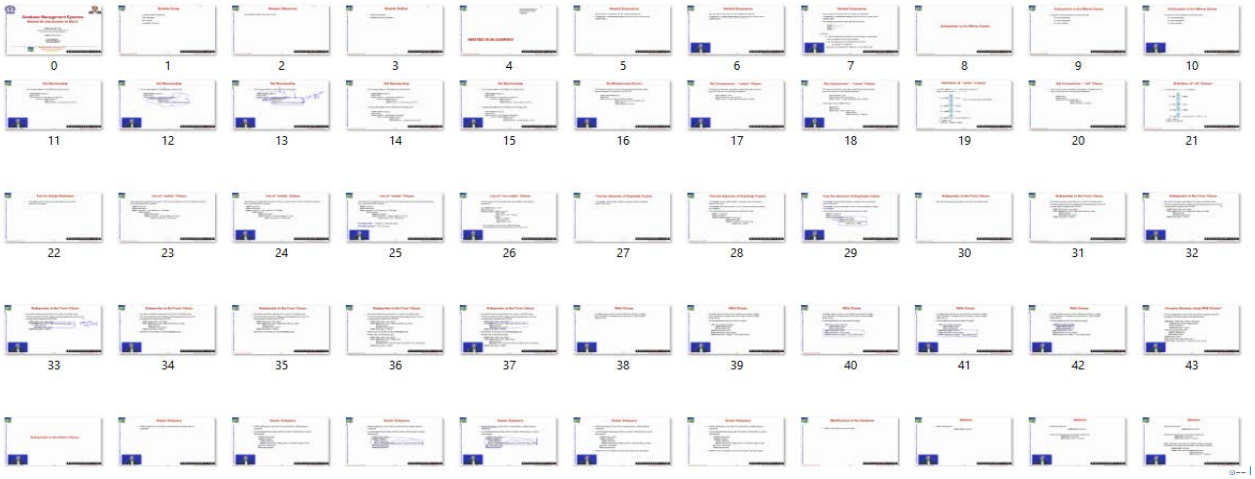


Fig. 3. Sample Result of Key-Frames.

One more asset of the GCV API is when doing a request for processing a picture; Google provides the power to imply the types of evaluation that must be on this picture. For instance, object detection, facial location, milestone recognition, and a lot more examination perform on the picture. The workflow of the Fig. 4 has been implemented using a python script.

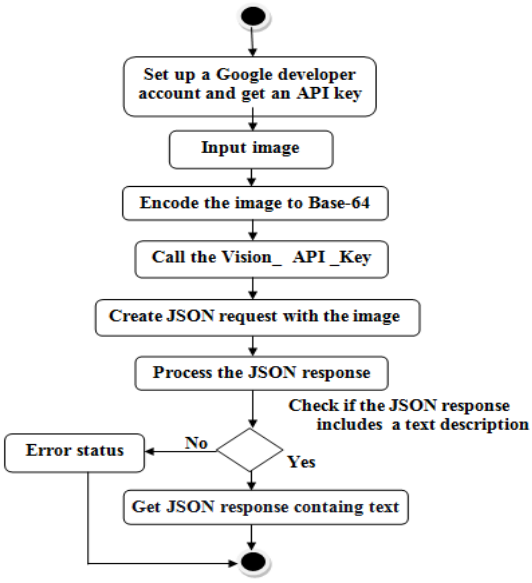


Fig. 4. Workflow of Image Text Extraction Process.

Google OCR has different advantages, here we depict the hugest advantages:

- Robust - The two capacities, serving two kinds of text records subject to the clients' choice, make the Google Vision OCR similarly more robust than single-model OCR tools.
- Language support - Google has exhorted that its OCR is appropriate to in excess of 60 languages.
- Ease of utilization - The actual model is important for the in-constructed Google
- Vision library. The function-calling technique can be utilized in various languages in an extremely clear way.
- Scalability - Google's evaluating technique promotes clients to increase the use of the API, as more use prompts a less expensive normal cost.
- Speed - Google Cloud's warehouse stage superbly goes with the API utilization. By transferring the pictures into the drive, the response or reaction time of API can be extremely quick and versatile.

a) API call: Indicate the URL to the API and include the JSON data to POST to it. We first need to set up a Google developer account and obtain an API key [7] to perform OCR using Google Cloud Vision API.

b) Request: Send a JSON request containing a base64 encoded image file. The vision API performs feature detection on an image file.

c) Response: We get a response in JSON format which includes text and bounding box containing location coordinates. Sample results of text extraction using Google Cloud Vision API is shown in Fig. 5.

2) Title Identification: For the most part, in the talk slides, the substance of title, caption, and main points has more importance than the typical slide text, as they sum up each slide. The design of text lines can mirror their diverse importance. This data is important for a talk video indexing. To recognize the potential title text lines, we apply the accompanying conditions.

a) The height of the title text line is more prominent than or equivalent to the normal height of text lines.

b) Title text line has, at any rate, three characters.

c) Horizontal start position of the text line ought to be, not exactly 50% of the frame width.

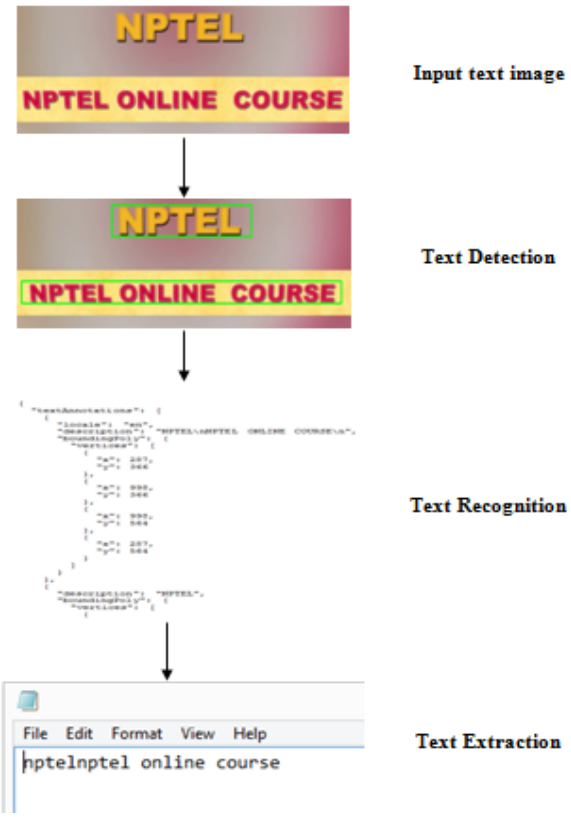


Fig. 5. Sample Result of Text Extraction from an Image.

C. Speech-to-Text (STT) Analysis

Text from speech is one of the principle wellsprings of data in a talk video. The teacher gives detailed data about the point in the video address. The speech text is generous and unconstrained. The speech is one of the significant variables in content-based recovery of a point in a long video address. Utilizing Google Speech-to-Text Programming interface as a speech recognition instrument in our trial, the speech records of talk recordings are used for indexing purposes. The speech text may differ marginally; the educator might talk some irregular substance. In any case, accepted that speech includes significant theme data and can be utilized to accomplish topic division and index point creation. Speech-to-text APIs provide lot of pros like boosting productivity and efficiency, saves time, Reliability, helps physically disabled people, etc. This API is used in several applications like Chatbots, Automated dictation, Smart assistant, Voice commanding, Transcriptions for call centers, mixed language detection, etc. Among the popular APIs like Microsoft Google Speech-To-Text, Cognitive Services, Dialogflow, IBM Watson, etc., for speech-to-text processing the Google Speech-To-Text API gives more accurate results.

3) Google Cloud Speech-to-Text API: The Google Cloud Speech API is integrated with Google Translate API and Cloud Vision API. Machine Learning is essential for the Google Cloud Stage in the development of applications that can listen, view, and comprehend its general surroundings. With this complete Google Cloud Speech API developers can easily translate an audio into text by using neural network models. This API supports 110 plus languages and variations, to help a worldwide user base. The Google Speech Programming interface, otherwise called Speech-to-Text (STT), is a modern instrument that uses Google's AI innovation to change voice over to message. Google Speech Programming interface is one of the most amazing speech recognition service [1]. This API is an automated speech recognition (ASR) API adapted with deep neural networks. It can likewise deal with noisy sound in a variety of conditions. This API result include not only text but also the timestamp corresponding to each word. The flowchart of Talk-To-Text (TTT) processing is shown in the Fig. 6.

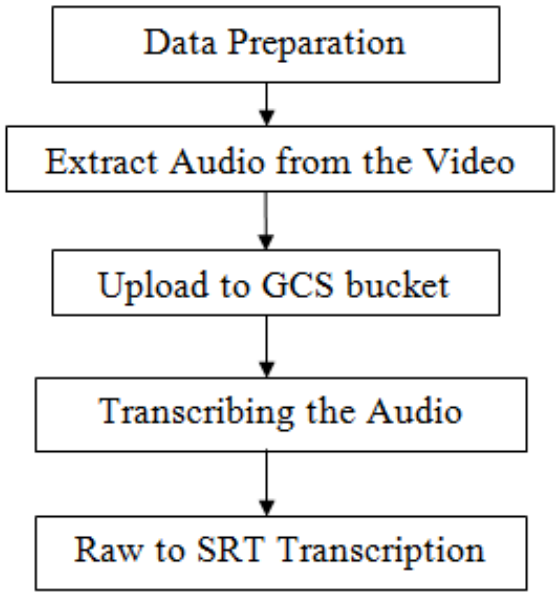


Fig. 6. Flowchart of TTT Processing.

a) Data preparation: Download lecture videos from different online courses like YouTube, NPTEL.

b) Extract audio from the video: Build the instance using Google speech-to-text API for Talk-To-Text (TTT) processing. Before doing anything, we have to install Ffmpeg to extract the audio from the lecture video. Here we are converting mp4 video file to ogg audio file. We have specified codec Opus in VoIP because of its audio compression with more quality and less delay rates. The sampling rate is set to 16000 hertz.

c) Upload to GCS bucket: We know that usually lecture video duration is longer than 60 seconds. Thus, we are requesting asynchronous speech recognition and we must store the audio file longer than 60 seconds in Google Cloud Storage bucket.

d) Transcribing the audio: The processing of audio file to obtain the transcription has been shown in the Fig. 7.

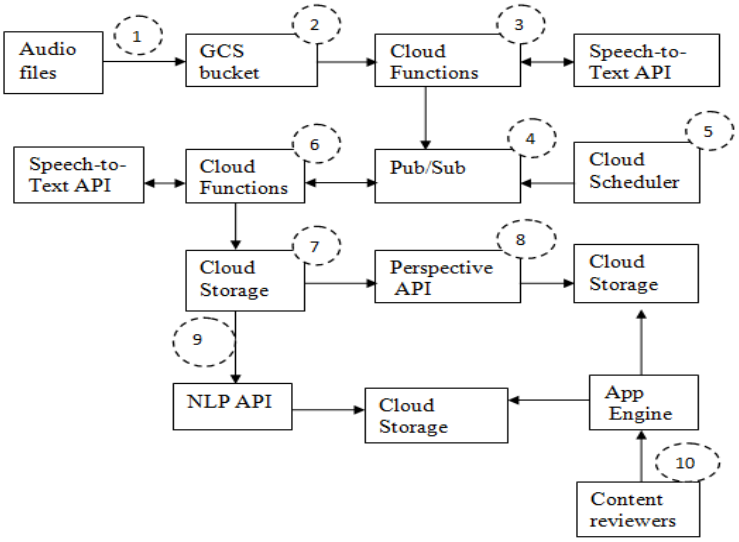


Fig. 7. Transcript Generation Process.

- Store audio file. The audio file is then stored in a Cloud Storage bucket. Before the audio file go through the remaining steps this step functions as a production bucket to maintain the files.
- Activate Cloud Function. When audio file meets the production bucket, a notification is generated. This notification triggers a Cloud Function to invoke a Speech-to-Text API.
- Invoke the Speech-to-Text API. Speech-to-Text API is invoked by Cloud Function to get a transcription of the audio file. This process is nonparallel, so a job ID is sent to the Cloud Function by Speech-to-Text API.
- Report Speech-to-Text job IDs. The audio filename and job IDs are then reported to the Pub/Sub point.
- Speech-to-Text voting. For every 10 minutes the Cloud programmer reports an announcement to a Pub/Sub point, which activates a next Cloud Function.
- Get Speech-to-Text API results. This Cloud Function extracts all announcements from the initial Pub/Sub point and pulls the filename and job IDs for every news. Each individual job status is checked by calling the Speech-to-Text API.
- In case when a job is over, the resulted transcription are registered to a next Cloud Storage bucket. Then Cloud Function moves the audio file to Cloud Storage bucket from the production Cloud Storage bucket. In case when a job is not over, a Pub/Sub announcement is added again to the Pub/Sub point. If there is no result from Speech-to-Text, the audio file is passed to a Cloud Storage error bucket. The obtained result (transcript of the audio file) is reported to a Cloud Storage bucket.
- Call Perspective API. The chance of "corruption" in the transcription is checked by calling Perspective API. Obtained results on this analysis is reported to another Cloud Storage bucket.
- Call the Cloud Natural Language API. The overall notion of the transcription is checked by calling the Natural Language API (called by fourth Cloud Function). The Cloud Function then reports the obtained results to another Cloud Storage bucket.
- Content reviewers. In the above diagram the App Engine enables user to check the outputs.

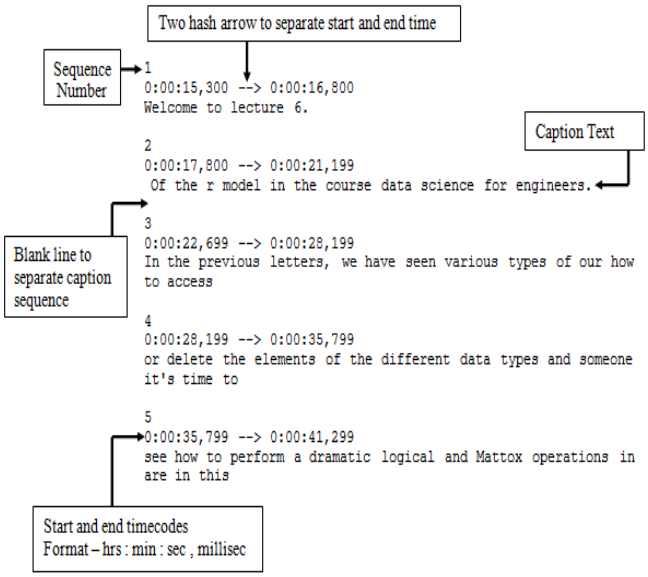


Fig. 8. SRT Formatted Text.

e) Raw to SRT Transcription: Raw text is the text obtained or collected from transcript generator before any manipulation. Thus, this text data is being send to the document pre-processor for analysing the raw text and produce resultant speech text with corresponding timestamps.

The raw text file is been formatted to SRT ("SubRip Subtitle") formatted text file because each sequence of text in .srt file has five important parts shown in the Fig. 8.

IV. Result and Discussion

The implementation is done in an Intel Core 8 CPU @ 5.0 GHz, with ubuntu operating system.

A. Keyframe Extraction

To evaluate the performance of video keyframe extraction, we randomly chose seven lectures videos like data science (DS), cryptography(crypt), cloud computing (CC), computer networks (CN), DBMS, algorithms (Alg), and machine learning (ML) from different online courses with varying layouts, font size, and styles.

TABLE I. Key-Frames

Lecture Video	Duration in minutes	Keyframes	Total slides
DS	22	38	49
ML	60	39	43
CN	23	38	42
crypt	100	75	77
DBMS	33	72	75
CC	45	51	57
Alg	109	125	147

The number of desired slides in the lecture videos are manually annotated for ground truth. Then, we applied the slide extraction algorithm to these videos. We compare the results of extracted slides with ground truth using recall and precision. The precision and recall esteem recognize bogus alert rate and missed location outline rate individually. The estimation of precision diminishes if there is over-segmentation, i.e., superfluous frames are separated. The assessment of Recall lowers if there is under- segmentation, i.e., an ideal frame stays undetected. The Table I shows the result of obtained keyframes.

The F1 score is measured as:

$$\text{F1 score} = \frac{(2 * \text{Precision} * \text{Recall})}{(\text{precision}+\text{Recall})} \quad (2)$$

Where,

$$\text{Precision} = \frac{\#\text{slides detected correctly}}{\#\text{slides detected}} \quad (3)$$

$$\text{Recall} = \frac{\#\text{slides detected correctly}}{\#\text{ground truth slides}} \quad (4)$$

Results of Precision, Recall, and F1 score obtained fromthe above formulas is 0.9, 0.98, and 0.94 respectively.

B. Slide-to-Text Conversion

With an average accuracy of 96.7 percent (Fig. 9), the text extraction results from each lecture video using the obtained key-frames showed that GCV outperformed other OCR APIs in this task. The findings are displayed in Table II. Transym's is 80.8 percent, Tesseract's is 92 percent, and Abbyy Finereader's is 90.5 percent. When the file size and resolution are taken into account, the accuracy of the GCV OCR is significantly higher than that of other methods. Additionally, low-resolution or small-size photos have the lowest accuracy. Performance is assessed using the three factors stated below.

$$Recall = \frac{Total\ Extracted\ text}{Total\ key-frames} \quad (5)$$

$$Precision = \frac{Correctly\ extracted\ text}{Total\ Extracted\ text} \quad (6)$$

$$F1 - score = equation\ (2)$$

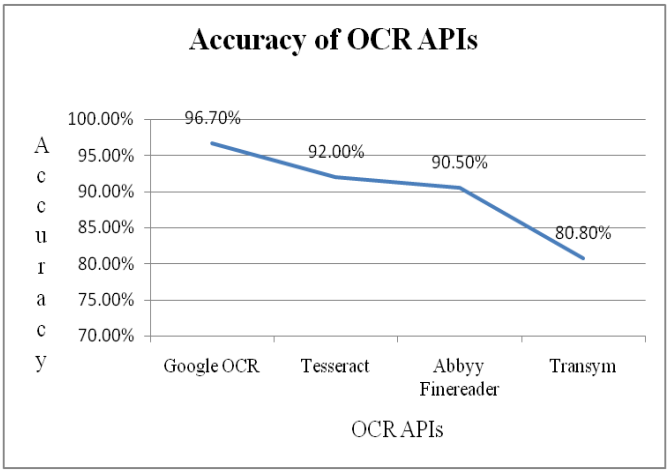


Fig. 9. Performance Comparison of Different OCR APIs.

C. Title Identification

Segmentation is done on lecture videos and taken 98 lecture slides to evaluate title identification. Then, the geometrical information of the text lines are used to identify the title line. The accuracy (Acc) of the title identification method is measured using the formula given below:

$$\text{Acc \%} = 1 - \left(\frac{\#\text{errors}}{\#\text{slides with title}}\right) * 100 \quad (7)$$

As a result, we obtained that the title line in 94 slides was identified correctly among 98 slides. The accuracy gained is 96%.

D. Transcript Generation

Three different speech to text APIs are used to perform this. The performance of each method is evaluated using the below three metrics

1) Word Error Rate (WER): It is used to test the occurrence of word errors in the obtained transcript. Levenshtein Distance is applied to find the difference between two word placements. As the word placement can have varying length, there can be substitutions (S), deletions (D), and insertions (I) to alter one word into the other. The WER can be calculated using the below equation (8).

$$\text{WER} = ((S+D+I)/N)*100 \quad (8)$$

TABLE II. RESULTS OF DIFFERENT OCR APIS

Method	Pr (%)	Re (%)	F1-Score (%)
Google OCR	97.2	94.7	**97.4**
Tesseract	88.2	89.4	88.7
Abbyy Finereader	87.8	86.8	87.2
Transym	65.6	84.2	73.7
Google OCR	94.7	97.4	**96.0**
Tesseract	86.1	92.3	89.0
Abbyy Finereader	91.4	89.7	90.5
Transym	72.7	84.6	78.1
Google OCR	97.2	97.3	**97.2**
Tesseract	88.8	94.7	91.6
Abbyy Finereader	88.2	89.4	88.7
Transym	62.5	84.2	71.7
Google OCR	97.2	97.3	**97.2**
Tesseract	91.5	94.6	93.0
Abbyy Finereader	91.3	92.0	91.6
Transym	83.3	88.0	85.5
Google OCR	98.5	98.6	**98.5**
Tesseract	92.6	94.4	93.4
Abbyy Finereader	95.5	93.0	94.2
Transym	86.1	90.2	88.1
Google OCR	98.3	96.0	**97.1**
Tesseract	93.4	90.1	91.7
Abbyy Finereader	88.6	86.2	87.3
Transym	75.6	80.3	77.8
Google OCR	89.4	98.4	**93.6**
Tesseract	96.6	96.8	96.6
Abbyy Finereader	94.1	95.2	94.6
Transym	88.8	93.6	91.1

2) Word Recognition Rate (WRR): Below equation (9) is used to calculate the WRR.

$$WRR = ((N-D-S) / N)*100 \quad (9)$$

3) Sentence Error Rate (SER): It is used to find the occurrence of errors in the sentences of the transcript. If there is a word by word match between manual transcription and recognition output, then it is taken into account as exact match. The SER accuracy is calculated using the below formula.

$$SER = \frac{\text{No.of inaccurate sentences}}{\text{Total No.of sentences}} * 100 \quad (10)$$

The results of WER, WRR, and SER obtained by three Speech-to-Text APIs is shown in Table III and it clearly shows that results of Google Speech-to-Text API is much more better than other two methods. IS stands for incorrect sentences. The comparison of three APIs (IBM, Microsoft, Google) in terms of WER, WRR, and SER is shown in Fig. 10.

TABLE III. RESULTS OF DIFFERENT SPEECH-TO-TEXT APIS

Google Cloud Vision OCR					
Words	***S***	***I***	***D***	***Sentences***	***IS***
2698	149	72	112	213	112
6904	349	123	294	548	288
3044	184	98	136	268	149
13922	773	388	528	1020	567
4620	238	118	178	358	192
7435	364	142	346	596	302
14222	788	391	593	1131	583
IBM Watson					
2530	189	122	142	205	133
6733	519	374	394	538	323
2874	272	128	186	259	186
13747	1298	949	1068	1009	589
4451	393	258	298	351	194
7267	614	392	436	587	349
14047	1328	992	1198	1119	595
Microsoft					
2488	197	184	188	198	132
6689	651	458	583	530	342
2826	255	203	218	248	181
13693	1543	1056	1284	1001	591
4404	468	282	346	341	198
7219	758	487	592	574	375
13997	1846	1083	1423	1113	602

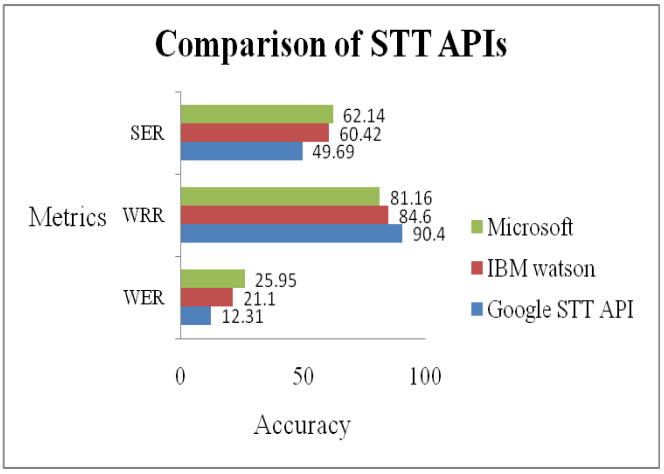


Fig. 10. Comparative Results of Three Different APIs.

V. CONCLUSION

This paper presents a whole work stream for keyframe extraction, hybrid text extraction and title identification proof. Frame differencing method is applied to accomplish superior key-frame extraction and achieved 94% of F1 score. To extract the text from key-frames four OCR methods have been proposed and found that Google cloud vision OCR is best achieving upto 97% accuracy. Google permits the API to handle singular bits of a picture independently and return the

outcome rapidly in brought together configuration. The title lines are identified using the geometrical information of the text lines and gained 96% accuracy. To extract the speech text three different APIs are used namely, Microsoft, IBM, and Google. The WER, WRR and SER are computed to measure the accuracy of these model and the achieved result of these parameters is shown in this paper. This paper founds that Google speech to text API has achieved best result in terms of WER, SER, and WRR compare to other two APIS. The outcomes acquired are really exact and very helpful.

The future work includes the implementation of an indexing algorithm for lecture videos based on obtained hybrid text.

ACKNOWLEDGMENT

We are very thankful to our parents, family, and friends for supporting to complete this work.

REFERENCES

[1] Foteini Filippidou and Lefteris Moussiades, "A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems," IFIP International Federation for Information Processing 2020 Published by Springer Nature Switzerland AG 2020. pp. 73–82, 2020. https://doi.org/10.1007/978-3-030-49161-1_7

[2] Yin J, Liu L, Liu Q. The infrared moving object detection and security detection related algorithms based on W4 and frame difference[J].Infrared Physics & Technology, 2016 (7), pp. 302 - 315.

[3] Akshay Parwar, Akansha Goverdhan, Apurva Gajbhiye, Prajkta Deshbhratar, Roshan Zamare, Prasanna Lohe, "Implementation to Extract Text from Different Images by Using Tesseract Algorithm," International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 6 Issue 2, Feb. 2017, pp. 20298-20300.

[4] Joshua Y. Kim1, Chunfeng Liu, Rafael A. Calvo, Kathryn McCabe, Silas C. R. Taylor, Björn W. Schuller, Kaihang Wu,. "A Comparison of Online Automatic Speech Recognition Systems and the Nonverbal Responses to Unintelligible Speech", 2019. https://doi.org/10.48550/arXiv.1904.12403

[5] Nurzam, F.D., Luthfi, E.T. "Implementation of real-time scanner java language text with mobile vision android based." In: 2018 International Conference on Information and Communications Technology (ICOIACT). IEEE; 2018, pp. 724–729.

[6] Veton Këpuska, Gamal Bohouta, "Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx)", Int. Journal of Engineering Research and Application, ISSN : 2248-9622, Vol. 7, Issue 3, (Part -2) March 2017, pp.20-24.

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, "Region-based convolutional networks for accurate object detection and segmentation," Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol. 38, no. 1, 2016, pp. 142–158.

[8] Shih-Hsin Chen, Yi-Hui Chen, "A New Content-Based Image Retrieval Method Based on the Google Cloud Vision API," ACIIDS 2017: Intelligent Information and Database Systems, 2017, pp 651-662.

[9] M. S. Barnish, D. Whibley, S. Horton, Z. R. Butterfint, and K. H. O. Deane, "Roles of cognitive status and intelligibility in everyday communication in people with Parkinsons disease: A systematic review," J. Parkinsons. Dis., vol. 6, no. 3, pp. 453–462, 2016.

[10] A. Behrman, "A clear speech approach to accent management," Am. J. speech-language Pathol., vol. 26, no. 4, pp. 1178–1192, 2017.

[11] Zhao, Baoquan, Songhua Xu, Shujin Lin, Ruomei Wang, and Xiaonan Luo. "A New Visual Interface for Searching and Navigating Slide-Based Lecture Videos." In 2019 IEEE International Conference on Multimedia and Expo, 2019, pp. 928-933,

[12] Huang, Cheng, and Hongmei Wang. "Novel key-frames selection framework for comprehensive video summarization," IEEE Transactions on Circuits and Systems for Video Technology (2019).

[13] A. M. Reddy, V. V. Krishna, L. Sumalatha and S. K. Niranjan, "Facial recognition based on straight angle fuzzy texture unit matrix," 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), Chirala, 2017, pp. 366- 372.

[14] Purushotham Reddy, M., Srinivasa Reddy, K., Lakshmi, L., Mallikarjuna Reddy, A. "Effective technique based on intensity huge saturation and standard variation for image fusion of satellite images" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5, June 2019.

[15] Wu, Jiaxin, Sheng-hua Zhong, Jianmin Jiang, and Yunyun Yang. "A novel clustering method for static video summarization." Multimedia Tools and Applications 76, no. 7 (2017), pp. 9625-9641.

[16] Zhang, Lanshan, Linhui Sun, Wendong Wang, and Ye Tian. "KaaS: A standard framework proposal on video skimming." IEEE Internet Computing 20, no. 4 (2016), pp. 54-59.

[17] Irmanti, D.,Hidayat, M.R., Amalina, N.V., Suryani, D., et al. "Mobile smart travelling application for indonesia tourism." Procedia computer science 2017, pp. 556–563.

[18] Chen, S.H., Chen, Y.H. "A content-based image retrieval method based on the google cloud vision API and wordnet." In: Asian Conference on Intelligent Information and Database Systems. Springer; 2017, pp. 651–662.

[19] Mulfari, D., Celesti, A., Fazio, M., Villari, M., Puliafito, A.. "Using google cloud vision in assistive technology scenarios." In: 2016 IEEE Symposium on Computers and Communication (ISCC). IEEE; 2016, pp. 214–219.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, 2015, pp. 436–444.

[21] D Intan, S Saputra, SW Handani, GA Diniary. Utilization of Cloud Speech API for the Development of English Language Learning Media using Speech Recognition Technology (in Indonesia Pemanfaatan Cloud Speech API untuk Pengembangan Media Pembelajaran Bahasa Inggris Menggunakan Teknologi Speech Recognition). TELEMATIKA. 2017; 10(2): 92–105.

[22] Hyun Jae Yoo, Sungwoong Seo, Sun Woo Im, and Gwang Yong Gim, "The Performance Evaluation of Continuous Speech Recognition Based on Korean Phonological Rules of Cloud-Based Speech Recognition Open API", International Journal of Networked and Distributed Computing Vol. 9(1); January (2021), pp. 10–18

[23] Foteini Filippidou and Lefteris Moussiades, "A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems", International Federation for Information Processing (IFIP), pp. 73–82, 2020.

[24] H. Roh and K. Lee, "A Basic Performance Evaluation of the Speech Recognition API ofStandard Language and Dialect using Google, Naver, and DaumKAKAO APIs", Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology, Vol.7, No.12, pp. 819-829, December 2017.

High Capacity Image Steganography System based on Multi-layer Security and LSB Exchanging Method

Rana Sami Hameed[1], Siti Salasiah Mokri[2]
Department of Electrical,
Electronic and Systems Engineering
Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia
Selangor, Malaysia

Mustafa Sabah Taha[3]
Missan Oil Training Institute, Ministry of Oil, Iraq

Mustafa Muneeb Taher[4]
College of Computing Science & Information Technology
University Tenaga Nasional, Selangor, Malaysia

***Abstract*—Data security is becoming an important issue because of the vast use of the Internet and data transfer from one place to another. Security of these data is essential, especially when these data represent critical information. There are several techniques used to hide these data, such as encryption. Steganography can be utilised as an alternative to encryption because encryption is susceptible to data modification during transmission. Steganography is the hiding data on a cover multimedia such as images, audio, and video. The technique allows security for data transmission so unwanted third parties cannot notice the hidden data. The challenge of steganography is the trade-off between the hidden data's payload capacity and the system's imperceptibility and robustness. If the hidden data increases, the imperceptibility and the robustness will be decreased. This case is a big challenge in this digital world where social media, Internet, and data transfer are used hugely. Because of this, this paper proposes using a modified Least Significant Bit (LSB) method for the embedding process called Multi-Layer Least Significant Bit Exchange Method (MLLSBEM). This proposed algorithm uses the AES encryption method to encrypt the secret text and then uses Huffman coding to compress the encrypted message as pre-processing data. The proposed study seeks to strike a compromise between important issues, provide maximum payload capacity, and retain high security, imperceptibility, and reliability for secret communication Using image processing and steganography techniques. Simulation findings demonstrate that the suggested method is superior for existing PSNR, SSIM, NCC, and payload capacity investigations. The proposed method is immune to the histogram, chi-square, and HVS attacks.**

Keywords*—Information hiding; steganography; cryptography; multi-layer security; high capacity component

I. Introduction

Increasing the use of data transfer from one place to another, especially with the improvement of the Internet and online transferring tools, the concern of data privacy and security becomes an important issue [1]. Third-party attacking of data is a risk. This circumstance means that attackers can reach the sender's data in many ways leading to the possibility of malicious threats, eavesdropping, and other malicious activities [2]. Three common techniques are used in this field to ensure data security and privacy: steganography, cryptography, and the watermarking technique. These three techniques are used to ensure security in different ways [3]. In steganography, the required data, which can be called the secret data that is to be sent, is hidden within media data such as images, video, audio, and protocol. On the other hand, in cryptography, the required data to be sent is coded with a predefined code known by both the transmitter and the receiver while in watermarking, the secret data does not change, but it is carried on multimedia data, such as images, to be sent to the required receiver to ensure confidentiality [4]. Image steganography is a widely used technique because the secret data does not change as in cryptography and does not appear as in watermarking. Image steganography can hide the secret data in its original format within the image by using a stego-key at the transmitter. The same stego-key is used at the receiver to recover the hidden data from the stego-image, as shown in Fig. 1.

Image steganography system has become the most popular research area than the other types of systems because of the availability, ease to use by users, and the ability to hold a large amount of data, in addition to the difficulty of noticing the hidden data by the unwanted third party [5].

However, using image steganography nowadays, where a high volume of data exchange is required, becomes a challenge, in particular to the issue of the payload capacity. Besides this, the trade-off challenge between payload capacity, security, and visual image quality is also a critical issue. These criteria imply that when the hidden data increases, the security of the steganography technique will decline and vice versa. Accordingly, as the amount of hidden data increases, the imperceptibility will decrease as well [6].

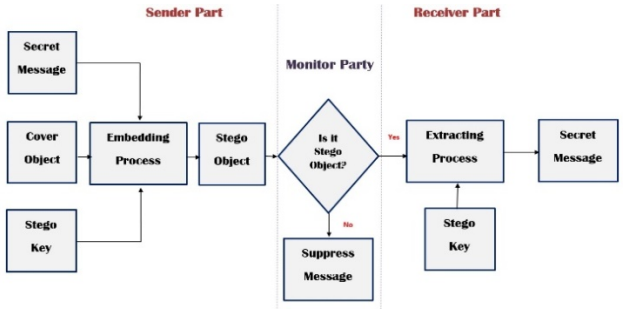


Fig. 1. The General Principle of the Image Steganography System.

Several methods and techniques have been used in the literature to overcome these challenges. The most straightforward and efficient method uses the Least Significant Bit (LSB) as an embedding technique, as in Gupta et al. [7] and Hameed et al. [8]. They propose LSB with some encryption techniques to ensure the system's security. Maurya & Gupta [9] suggested a method based on the adaptive LSB substitution steganography technique. In this method, the aim is to divide the image into two segments - non-sensitive and sensitive parts based on texture analysis. The majority of the bits in the non-sensitive area are used to hold secret messages and other bits in the sensitive area. The main advantages of this method are achieving both high payload capacity and imperceptibility.

The challenge of using LSB is the payload capacity of the system. Nowadays, most researchers, according to the best of our knowledge, use Artificial Intelligence (AI) and Deep Learning (DL) algorithms to overcome the capacity challenge. The study by [10] proposed the General Adversarial Networks (GAN) as a form of DL algorithm. GAN employs game theory to train the generative model using an adversarial method based on two networks for generator and discriminator. The data is fed into the generator model, and the result is approximately equal to the input image. The discriminator networks determine the class of the created images. Using GAN in steganography allows for increasing the payload capacity of the system, but this approach increases the complexity of the developed steganography technique. This complexity arose from the training data requirement before applying steganography to train the model. To overcome this, Volkhonskiy et al. [11, 12] proposed a Steganographic GAN (SGAN) based on DCGAN to simplify the training process as in Shi et al. [13] and [14]. They propose four fractionally convolutional layers followed by a functional layer with Hyperbolic tangent activation with base WGAN. The proposed model is based on multi-layer steganography but also depends on eavesdropping on the generator. Simulation results show that using four layers increases the receiver's prediction complexity, so Yang et al. [15] and [16] use the SGAN with three layers process. They use pixel-wise segmentation. This use reduces the complexity of the process but increases the discriminator and generator losses.

The work in [17] uses fuzzy logic to make decisions based on local statistical, texture, and brightness information-based feature vectors. The fuzzy logic can be used instead of AI algorithms to reduce the decision complexity. Simulation results show that at higher embedding rates, the approach helps to eliminate stego-image distortions. The study by [18] explained another fuzzy-based technique in which, before the real embedding process, the cover pixel selection is based on fuzzy pixel classification, and the secret message is translated to a mode of fuzzy data.

Fuzzy logic can improve stenographic techniques in various ways, especially when there is vagueness in the image textures. It benefits the system by recognising appropriate visual patterns quickly and avoiding irreversible complexities. However, adding a fuzzy process to select the embedding process affects the system's capacity.

This paper proposed a Multi-Layer Least Significant Bit Exchange Method (MLLSBEM) algorithm as an effective hiding method to embed a high volume of security data into an image without affecting the security and imperceptibility of the system. Thus, several contributions of this paper can be summarised as the following:

- This paper proposes a hiding method based on multi-layer operations, which is Advanced Encryption System (AES) technique to enhance security. Furthermore, the paper uses Huffman coding to increase the payload capacity of the hidden data.
- The propose method is based on Linear Significant Bit Exchanging Method (LSBEM) that is a simple, secure, and efficient method used with image steganography based on the resulting image quality.
- The evaluation of the proposed multi-layer algorithm is performed on grey-scale and colour images to generalise its use and efficiency.

The remaining of this paper is structured as follows: the proposed algorithm description is discussed in Section II. The overall methodology and simulation parameters are mentioned in Section III. Simulation results and discussion based on performance metrics are presented in Section IV. Finally, the conclusion and some future work discussions are mentioned in Section V.

II. The Proposed Algorithm

This newly proposed technique aims to have a stego-image like the original image with high capacity hidden data. The aim is to satisfy the imperceptibility of maintaining a high PSNR value. The proposed MLLSBEM algorithm depends on two methods: designing a new LSBEM and using Implicit Key Generation (IKG). The proposed LSBEM technique's principle is to expand the LSB substitution method to exchange the secret message bits with the cover image's pixels using the shared IKG between the sender and receiver.

Algorithm 1: The proposed hiding algorithm

If STB1 = CIPLpixel,

The BL of the pixel is set to ‘1'; otherwise, it is set to ‘0'.

If STB2 = CIPRpixel,

The BR of the pixel is set to ‘1'; otherwise, it replaces it with a ‘0'.

If either BL or BR = 0,

The next cover pixel's 2-LSBs are substituted with the previously unmatched secret bit pair (i.e., STB1 or STB2).

If BL and BR = 0,

The skipped mapped block. “there is no mapping”.

Algorithm 1 expresses the proposed algorithm to hide the secret payload in the carrier image. While Fig. 2 depicts the proposed scheme of the concealing process. As shown, the Secret Text Bits (STB) are initially grouped into two secret bit pairs, i.e. STB1 (1,0), STB2 (1,0), and STB3(1,1), as shown in Fig. 2(i). Similarly, all the cover image pixels bits are divided into pairs, as shown in Fig. 2(ii), with Left Pair called Cover Image Left Pixel (CIPLpixel) and the Right Pair called Cover

Image Right Pixel (CIPRpixel). CIPLpixel, pixel denotes the cover pixel's 7th and 8th Most Significant Bits (MSBs), while CIPRpixel denotes the 5th and 6th MSBs.

The pixel on the cover of the secret text bits pair would be mapped using CIPLpixel, and CIRpixel bits. Similarly, the 1st and 2nd LSBs of a cover pixel are represented by Right Bit (1LSBp) and Left Bit (2 LSBp), respectively, and are utilised as indicators for pixel pair mapping. In terms of indicator bits, 2LSBp, and 1LSBp, the pixel is closely related to CIPLpixel, and CIPRpixel, respectively, as shown in Fig. 2(iii).

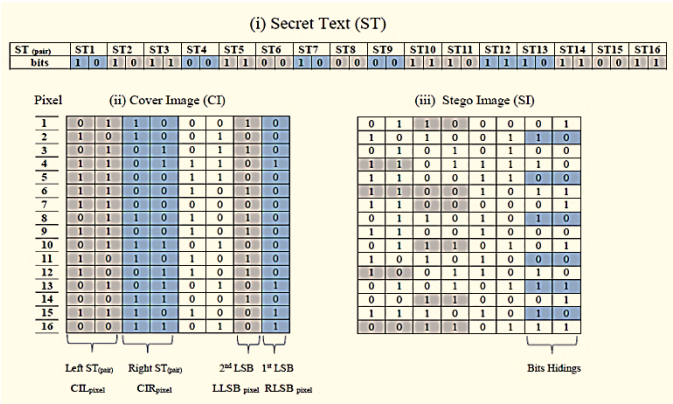


Fig. 2. The Full Scenario of the Proposed Embedding Process: (i) Secret Text Bits; (ii) Cover Image Pixels; (iii) Stego-image Pixels.

III. System Model and Methodology

Four phases describe the whole methodology of this paper, as shown in Fig. 3. These four phases are:

- Phase one: Data pre-processing. There are two stages in this phase which are (i) the secret message pre-processing and (ii) cover image preparation. In secret message processing, AES and Huffman coding are used while image normalisation is performed in the cover image preparation stage.
- Phase two: Embedding process. This phase also has two stages which are (i) LSB exchanging method and (ii) IKG, as described in Section II.
- Phase three: Evaluation process. This phase uses objectives and subjective evaluation metrics to test and evalaute the stego-image that includes the secret message. Some of the metrics used are Peak Signal-to-Noise Ratio (PSNR), Normalized Cross-Correlation (NCC), Structural Similarity Index Metric (SSIM), Mean Square Error (MSE), and payload capacity.
- Phase four: Extraction process. The objective of the extracting process is to extract the embedded data (secret bits) from the LSB pixels and simultaneously follow the procedure designed in the embedding process. Most of the information related to the extracting stage is made by the agreement between the sender and receiver parties.

The proposed methodology starts by applying data encryption using AES to convert the secret readable message to a non-readable message, raising the redundant characters, and then Huffman coding is used to reduce the size of the redundant character as much as possible. The proposed system perfectly deals with the USC-SIPI image database from the University of Southern California with different image sizes [15]. To work with this dataset, a preprocessng stage is needed to select the target pixel from the carrier image called "image normalisation process" before the embedding process. The proposed embedding algorithm (Stage 2), is applied. Then, the extraction process starts. The secret text bits are extracted by referring to the indicator bits (BLpixel or BRpixel) that the 2-LSBs substituted in each pixel. The extracting strategies are illustrated in Algorithm 2.

Algorithm 2. The proposed extracting algorithm.

First Action: If BL, the pixel is '1' and BR, the pixel is '0':
Restore CILpixel as STB1, and
Restore the 2-LSBs as STB2.

Second Action: If BL, the pixel is '0' and BR, the pixel is '1':
Restore CIRpixel as STB2, and
Restore the 2-LSBs as STB1.

Third Action: If both BLpixel, and BRpixel is '1', then
Restore STB1 from CILpixel bits, and
Restore STB2 from the CIRpixel bits.

Fourth Action: If both BLpixel, and BRpixel are '0', then
No mapping indication.

Table I shows the simulation parameters used in this paper. The image size used is 512×512 with coloured and grey-scale images type. The payload capacities are 16384, 32768, and 49152 (16 kB, 32 kB and 49 kB). The PSNR threshold is 30 dB to consider imperceptibility.

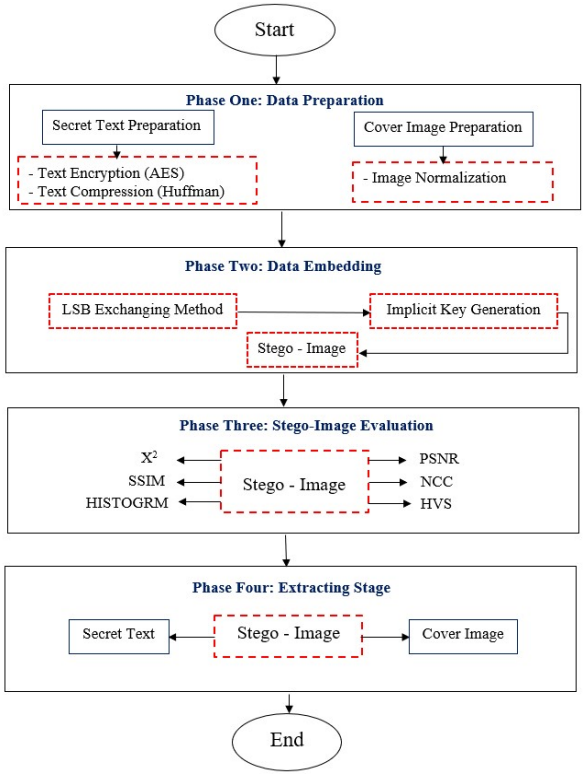


Fig. 3. The General Flowchart of the Proposed Methodology.

TABLE I. SIMULATION PARAMETERS

Simulation parameter	Values
Image size	512×512
Image format	TIFF format
Database used	USC-SIPI
Normalization	No
Payload capacity	16384, 32768, and 49152
PSNR threshold	30 dB
Coding	Huffman
Encryption	AES

IV. SIMULATION RESULTS AND DISCUSSION

In steganography, the imperceptibility and capacity of the proposed algorithm are the main concern in the simulation results. As known, there is a trade-off between capacity and imperceptibility of steganography. The proposed algorithm is evaluated based on the Embedding Capacity (EC) and three attacks system: the Human Visual System (HVS) attack, the Chi-square (χ^2)attack, and the Histogram attack. The comparison between the proposed algorithm with previously proposed methods in the literature is also presented.

The proposed algorithm uses colour and grey-scale images with different payload capacities which are 16 kB, 32 kB, and 49 kB. The 16384 bytes corresponded to 6.25%, meaning that every two pixels represented 16 bits; thus, 1/16 = 6.25% when 1 bit of two pixels was embedded. The 32768 bytes were equal to 12.5%, implying that every pixel corresponded to 8 bits, so 1/8 = 12.5% when 1 bit of one pixel was embedded. The 49152 bytes corresponded to 18.75%, signifying that every two pixels were assigned to 16 bits; accordingly, 3/16 = 18.75% when 1.5 bits of one pixel were embedded.

The PSNR threshold should be $\geqslant$ 30 dB in evaluating the imperceptibility to satisfy no HVS. In conventional image processing, the imperceptibility of the stego image is determined using the PSNR measures. By applying the PSNR measures, the fidelity of the stego image can be evaluated against the original carrier image. To ensure that the proposed system achieves the target aim, the proposed MLLSBEM is evaluated with two other embedding algorithms, the simple LSB and the embedding with pre-processing. Moreover, the proposed system is evaluated using various payload capacities: 16 kB, 32 kB, and 49 kB and three colour and grayscale images: Lena, Baboon, and Pepper. Table II presents the achievements of different embedding methods of grayscale images with 16 kB of payload capacities.

Simulation results in Table II show that the best PSNR comes from the proposed MLLSBEM because of the AES and Huffman coding compared to PSNR value of with pre-processing and simple LSB. Furthermore, the PSNR factor took the frequency of bits in the LSB of the cover image. The same conclusion comes from all the images via high PSNR value and low MSE. Fig. 4 shows the results of the proposed MLLSBEM using various payload capacities and various USC-SIPI dataset images Lina, Baboon and pepper images.

Coloured images are also used to evaluate their performance for the same payload capacities used in grey-scale images. Fig. 5 shows the PSNR values for the proposed algorithm when using coloured images. Simulation results show that the calculated PSNR values for the colour images are lower than the grey-scale images due to the representation of colour pixels with 24-bits for one pixel as opposed to only 8-bits for the grey scale. In addition, the Baboon image showed a higher PSNR value due to the different properties of this image, and also the nature of the image itself has more contrasts in the pixel value, thereby enabling the Baboon image to be more chaotic.

The proposed method used different evaluation metrics such as MSE, SSIM, and NCC to check the stego image before sending it to the authorised receiver to ensure that familiar attacks such as HVS, Chi-square and Histogram are unable to detect the secret message. Table III shows the various evaluation metrics with different amounts of embedding capacity for the standard grey-scale and coloured images.

TABLE II. THE DIFFERENT EMBEDDING METHODS OF GRAY-SCALE WITH 16 KB OF PAYLOAD CAPACITY

Image/ Dataset	PSNR evaluatiom metric		
	Simple LSB	***With Pre-processing***	***MLLSBEM (Proposed)***
Lena	65.225	71.332	78.094
Baboon	66.331	72.099	78.129
Pepper	64.997	71.009	78.1491

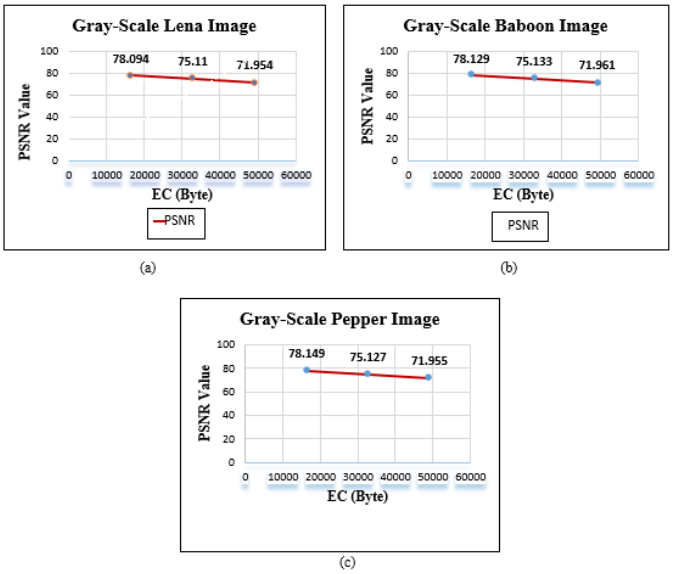


Fig. 4. The PSNR Values of Grey-scale Images with Various Payload Capacities. (a) Lena Image, (b) Baboon Image, (c) Pepper Image.

TABLE III. VARIOUS EVALUATION METRICS OF GRAY-SCALE AND COLORED IMAGES

Images	Embedding Capacity	16384 Bytes			32768 Bytes			49152 Bytes		
	Metrics	*MSE*	*SSIM*	*NCC*	*MSE*	*SSIM*	*NCC*	*MSE*	*SSIM*	*NCC*
Lena	Gray-Scale	0.0132	1	1	0.0144	0.998	0.976	0.0188	0.987	0.959
	Colored	0.0146	1	1	0.0146	0.998	0.987	0.0177	0.988	0.978
Baboon	Gray-Scale	0.0145	1	1	0.0149	0.999	0.997	0.0198	0.986	0.978
	Colored	0.0152	1	1	0.0159	0.999	0.997	0.0189	0.989	0.968
Pepper	Gray-Scale	0.0136	1	1	0.0175	0.998	0.998	0.0210	0.979	0.989
	Colored	0.0149	1	0.999	0.0169	0.997	0.989	0.0222	0.977	0.989

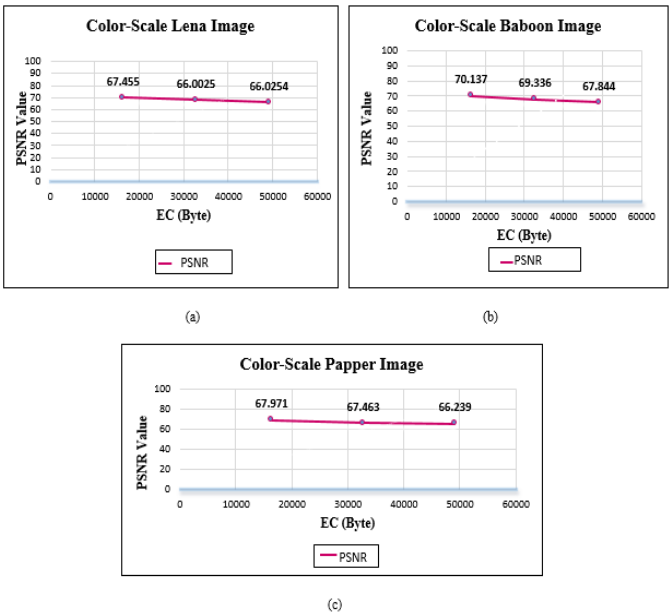


Fig. 5. The PSNR Values of Coloured Images with Various EC Values. (a) Lena Image, (b) Baboon Image, (c) Pepper Image.

A. Human Visual System Attacks Analysis

In image steganography, a system can detect edges that become blurred or unclear. Consequently, the LSB can detect the HVS attack, which is still ambiguous to human sight because it is trained to recognise the known things. This simulation aims to distinguish the presence or absence of the hidden data in the stego-image. Simulation results in Fig. 6 show that the eight-bit planes HVS attack can detect only the LSBs; the rest are ignored. The embedding in the bit planes 1 and 2 appears very clear where the vertical lines refer to the frequencies in their bits, implying that hidden information is embedded in these two-pixel. This type of detection is somewhat interactive between the system and humans because the system generates the pattern, and the human eyes detect it.

Fig. 7 depicts the comparison of the proposed MLLSBEM towards HVS attack as compared to the simple LSB and with pre-processing. Observations indicate that embedding simple LSB images is ineffective because the injection of the hidden bits directly, without processing or bit position selection, makes them immediately detectable by human eyes.

Due to preparation for the usage of picture normalisation, the pattern of the first bit-plane was improved for the pre-processing procedure. Using AES encryption and Huffman coding, the proposed MLLSBEM can generate a stego-image similar to the original cover image due to the arbitrary distribution of bits and the process of preserving the original image bit values by mapping the secret bits before embedding them.

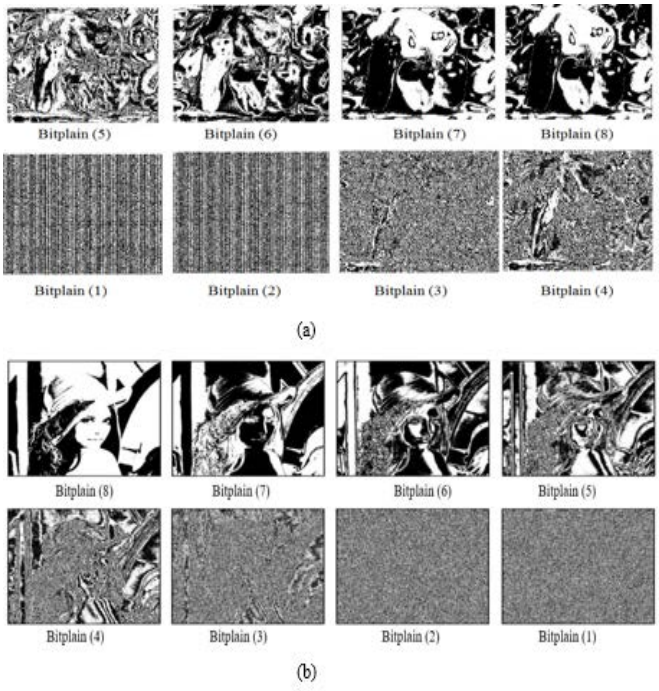


Fig. 6. The HVS Attacks with Various the Eight Bit-plains Layers. (a) Pepper Image, (b) Lena Image.

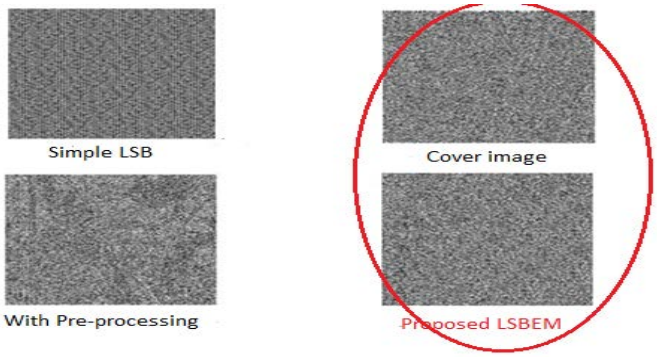


Fig. 7. The Figure shows Three Embedding Methods in Resistance against the HVS Attack on the Original Lena Image.

B. Chi-square Attack (X^2) Attack

A special attack called Chi-square (χ2) is based on the statistical analysis of the Pairs of Values (PoVs) exchanged during the secret data embedding, which is also based on the probability distribution. The (χ2) attack can find the probability of embedding the secret bits inside the stego image where the normal image follows the usual behaviour.

Fig. 8(a) shows the χ2-test for the original pepper image. In the first 10%, the probability is 0.065 because when the function checks the pixels, most of the characters in the alphabet start with the same value as the frequent bits. Thus, the test detects this frequently and suggests these pixels as the embedded data. The absence of detection for embedding in the remaining images is normal, as the original images do not have any hidden data.

Fig. 8(b) shows the χ2-test for the simple LSB that detects fifty per cent of the image as concealed data with a probability of one. In Fig. 8(c), the proposed algorithm covers the entire image with a low probability, even better than the original image. This occurs because the statistical distributions of the values in the LSB are good. After all, the segments of the secret bits are chosen carefully.

C. Histogram Attack

The histogram analysis is applied to three types of images, as shown in Fig. 9. The analysis shows that the variance between the constructed histograms is comparatively low for all tested images when using the proposed MLLSBEM method. The distortions caused by the embedding process are not noticeable to the human eye when concealing an acceptable amount of secret bit. However, when we hide more secret bits (exceeding the embedding limit), the variance between the constructed histograms is high and noticeable to the human eye. Moreover, the PSNR value becomes low.

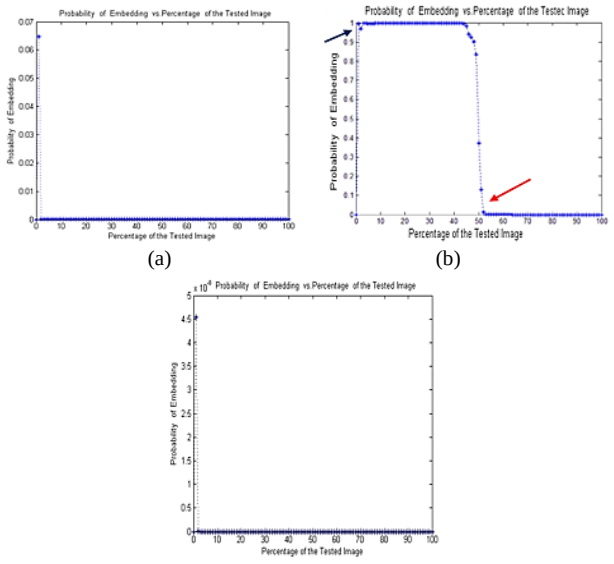


Fig. 8. The χ2-test for Pepper Image. (a) The Original Image, (b) The Simple LSB, and (c) The Proposed MLLSBEM for 16 kB Pixels.

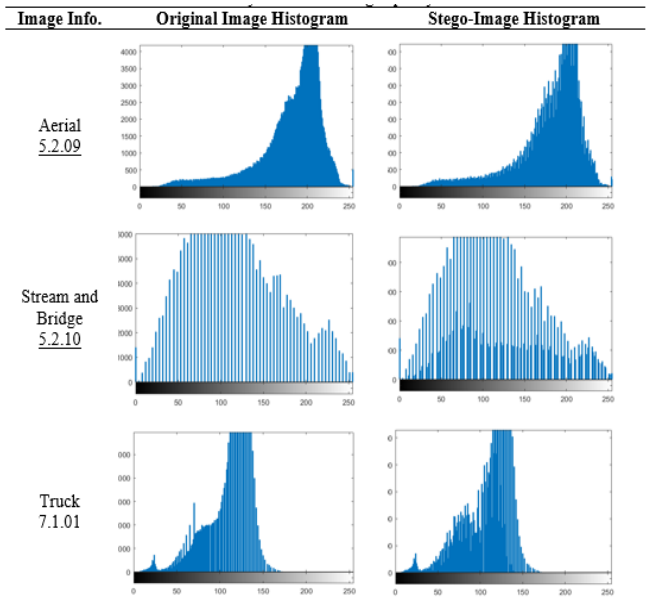


Fig. 9. Histogram Analysis of 16 kB Capacity of Grey-scale Image.

Fig. 10 shows the comparison between the stego images and the original image with different payload capacities. 18.75% embedding corresponds to 49 kB payload capacity.

Fig. 10. The Stego and Original Images' Resemblance with Different Payload Capacity.

Table IV compares the proposed algorithm's results with the previously published methods. The evaluation results of the proposed method were found to be better than those reported in the literature. This evaluation indicates that the proposed algorithm used for the pre-processing and embedding stages in the proposed scheme improve the final results. The proposed method is excellent as compared to the previous studies in terms of PSNR, SSIM, NCC, MSE for different payload capacities. In addition, the proposed method is resistant towards the histogram, chi-square, and HVS attacks.

TABLE IV. THE EVALUATION RESULTS OF THE PROPOSED METHOD

Cover Image 512 X 512 Colour	**Lena**					**Baboon**					**Pepper**				
	Capacity (KB)	***PSNR***	***SSIM***	***NCC***	***MSE***	***Capacity (KB)***	***PSNR***	***SSIM***	***NCC***	***MSE***	***Capacity (KB)***	***PSNR***	***SSIM***	***NCC***	***MSE***
A. S. Ansari et al. 2017 [19]	34	59.19	-	-	1.93	34	59.21	-	-	1.90	34	58.95	-	-	2.02
S.A. Parah et al., 2018 [20]	16	55.80	-	0.99	-	-	-	-	-	-	16	52.88	-	0.99	-
R. Shanthakumari and S. Malliga, 2019 [21]	78	54.85	0.99	1	0.85	78	54.62	0.99	1	0.90	78	54.73	0.98	1	0.88
A.S. Ansari et al., 2020 [22]	35	66.67	-	-	-	35	69.45	-	-	-	35	67.16	-	-	-
L. Tang et al. 2021 [23]	80	65.7	0.99	-	-	56	66.9	1	-	-	80	65.4	0.99	-	-
Proposed Method	49	66.02	0.98	0.97	0.0177	49	67.84	0.98	0.96	0.018	49	66.23	0.97	0.98	0.02

V. CONCLUSION

We have presented a new image steganographic method using the MLLSBEM to hide the secret bits. The proposed algorithm uses pre-processing operations such as AES encryption and Huffman coding. The proposed embedding algorithm is based on the modified LSB exchanging method and IKG. Simulation results show that the proposed method is excellent as shown in the achievement of PSNR, SSIM, NCC, and payload capacity and is robust towards the histogram, chi-square, and HVS attacks. In the future, we plan to take into consideration an open challenge to achieve adaptable exchange between the cover image and secret bits. This work opens up several new avenues that are worth doing for the future. For instance, the security can be enhanced by mixing the frequency domain and spatial domain. This may achieve better results in terms of security and robustness. Also, the proposed method can be combined with the DWT and embedding may result in high coefficients based on the obtained findings. Many methods have already used high coefficients for the embedding. However, the use of LSBEM may yield better results in terms of security and imperceptibility. The most important gap in the steganography system is related to the capacity improvement of the secret message. The limitation of the secret message with PSNR makes the steganography difficult to improve. In such a scenario, it is better to handle the secret message before embedding and to make it dynamic with the embedding method. The concealed message's limitation (direct hiding) makes steganography difficult to improve. In this case, it's preferable to manipulate the secret message before embedding it by coding and compressing it, making the secret text pre-processing stage interactive with the embedding process.

ACKNOWLEDGMENT

The authors would like to acknowledge Universiti Kebangsaan Malaysia and the Ministry of Education, Malaysia (MOE) for the Research University Grant with code: GUP-2019-023 to support this project.

REFERENCES

[1] S. Dhawan, C. Chakraborty, J. Frnda, R. Gupta, A. K. Rana, and S. K. Pani, "SSII: secured and high-quality steganography using intelligent hybrid optimisation algorithms for IoT," *IEEE Access*, vol. 9, pp. 87563-87578, 2021.

[2] Taha, Mustafa Sabah, et al. "High payload image steganography scheme with minimum distortion based on distinction grade value method." *Multimedia Tools and Applications* (2022): 1-34.

[3] X. Duan *et al*., "High-capacity image steganography based on improved FC-DenseNet," *IEEE Access*, vol. 8, pp. 170174-170182, 2020.

[4] E. Emad, A. Safey, A. Refaat, Z. Osama, E. Sayed, and E. Mohamed, "A secure image steganography algorithm based on least significant bit and integer wavelet transform," *Journal of Systems Engineering and Electronics*, vol. 29, no. 3, pp. 639-649, 2018.

[5] Hadad, Abbas Abd-Alhussein, et al. "A Robust Color Image Watermarking Scheme Based on Discrete Wavelet Transform Domain and Discrete Slantlet Transform Technique." *Journal homepage: http://iieta. org/journals/isi* 27.2 (2022): 313-319.

[6] Y. Ren, T. Liu, L. Zhai, and L. Wang, "Hiding Data in Colors: Secure and Lossless Deep Image Steganography via Conditional Invertible Neural Networks," *arXiv preprint arXiv:2201.07444*, 2022.

[7] A. Gupta, H. Shukla, and M. Gupta, "A Secure Image Steganography using X86 Assembly LSB," *NEU Journal for Artificial Intelligence and Internet of Things*, vol. 1, no. 1, pp. 38-47, 2022.

[8] M. A. Hameed, M. Hassaballah, S. Aly, and A. I. Awad, "An adaptive image steganography method based on histogram of oriented gradient and PVD-LSB techniques," *IEEE Access*, vol. 7, pp. 185189-185204, 2019.

[9] I. Maurya, and S. K. Gupta. "Secure image steganography through pre-processing." *In Soft Computing: Theories and Applications*, pp. 133-145. Springer, Singapore, 2019.

[10] S. Venkatesh, V. Sivakumar, A. K. Vagheesan, S. Sakthivelan, K. J. Kumar, and K. K. Nagarajan. "GANash -- A GAN approach to steganography." arXiv preprint arXiv:2110.13650, 2021.

[11] D. Volkhonskiy, B. Borisenko, and E. Burnaev. "Generative adversarial networks for image steganography". *Open Review*, 2016.

[12] D. Volkhonskiy, I. Nazarov, and E. Burnaev. "Steganographicgenerative adversarial networks." *In Twelfth International Conference on Machine Vision (ICMV 2019)*, vol. 11433, pp. 114333M. International Society for Optics and Photonics, 2020.

[13] H. Shi, X. Zhang, S. Wang, G. Fu, and J. Tang. "Synchronised detection and recovery of steganographic messages with ad-versarial learning." *In International Conference on Computational Science*, pages 31–43. Springer, 2019.

[14] H. Shi, J. Dong, W. Wang, Y. Qian, and Xiaoyu Zhang. “Ssgan: secure steganography based on generative adversarial networks.” *In Pacific Rim Conference on Multimedia*, pages 534–544. Springer, 2017.

[15] J. Yang, K. Liu, X. Kang, E. K. Wong, and Y. QingShi. “Spatial image steganography based on generative adversarial network.” *arXiv preprint*, arXiv:1804.07939, 2018.

[16] J. Yang, D. Ruan, J. Huang, X. Kang, and Y. QingShi. “An embedding cost learning framework using gan.” *IEEE Transactions on Information Forensics and Security*, 15:839–851, 2019.

[17] H. Dadgostar, and F. Afsari. “Image steganography based on interval-valued intuitionistic fuzzy edge detection and modified LSB.” *Journal of Information Security and Applications*, 30: 94–104, 2016.

[18] M. Islam, A. Roy, and R. Laskar. “Neural network based robust image watermarking technique in LWT domain.” *Journal of Intelligent & Fuzzy Systems*, 34(3): 1691–1700, 2018.

[19] A.S. Ansari, M.S. Mohammadi, and M.T. Parvez. “JPEG Image Steganography based on Coefficients Selection and Partition.” *International Journal of Image, Graphics and Signal Processing*, 9(6), 14, 2017.

[20] S.A. Parah, J.A. Sheikh, J.A. Akhoon, N. A. Loan, and G.M. Bhat, “ Information hiding in edges: A high capacity information hiding technique using hybrid edge detection.” *Multimedia Tools and Applications.* 77(1): 185–207, 2018.

[21] R. Shanthakumari, and S. Malliga. “Dual-layer security of image steganography based on IDEA and LSBG algorithm in the cloud environment.” *Sadhana - Academy Proceedings in Engineering Sciences*, 44(5), 2019.

[22] A.S. Ansari, M.S. Mohammadi, and M.T. Parvez. 2020. “A multiple-format steganography algorithm for color images.” *IEEE Access*, 8: 83926–83939S, 2020.

[23] L. Tang, D. Wu, H. Wang, M. Chen, and J. Xie. “An adaptive fuzzy inference approach for color image steganography.” *Soft Computing*, 25(16): 10987–11004, 2021.

Recognition of Odia Character in an Image by Dividing the Image into Four Quadrants

Aradhana Kar, Sateesh Kumar Pradhan
Department of Computer Science & Applications
Utkal University
Bhubaneswar, Odisha, India

***Abstract*—This paper deals with optical character recognition of Odia characters written in a particular font family 'AkrutiOriAshok-99' with different font sizes 18, 20, 22, 24, 26, 28, 36, 48 and 72 in Bold style. The font 'AkrutiOriAshok-99' is a font from the typing software 'Akruti'. The basic idea behind the approach followed in this paper is the character decomposition into four quadrants and then extracting features from each quadrant. The image processing techniques like converting the image to gray, resizing of image and converting gray image to binary are used in this approach. The system explained in this paper has two major parts: DictionaryBuilding and FindingMatch. For DictionaryBuilding, dictionary of images which are created either by scanning a document or a document converted to image, both written in same font family in different sizes. The features are extracted from each image in any font size in the 'Dictionary' using Preprocessing, FindPath, GettingFeaturesLeft or GettingFeaturesRight, VisitSubQuad, RemainingSubQuad, WriteToExcel and CommonFeature modules. The part FindingMatch is responsible for finding a correct match in the dictionary for the input image. For this, FeatureExtraction and Recognition modules have been used. Longest Common Subsequence (LCS) has been used for finding the common feature in DictionaryBuilding as well as finding the correct match. A total of 1800 characters, 200 characters of each font size have been tested and 98.1% of correctness has been achieved.**

***Keywords*—Odia characters; image processing; character decomposition; machine learning; optical character recognition**

I. INTRODUCTION

In present days, the textual data are either scanned or converted to image by using software to store the data in the form of image. It is required to recognise the characters present in the scanned document or document converted to image by using some algorithm. For recognition of characters in an image, efficiency in the segmentation of lines, words and characters should be achieved.

In Odia language, the alphabets are grouped into three categories: Swara Barna, Byanjana Barna and Atirikta Barna [1] (Fig.1). Only Chandra Bindu ( ), Anusara ( ) and Bisarga ( ), which are part of Byanjana Barna can be used with all the alphabets of Swara Barna, other alphabets of Byanjana Barna and all alphabets of Atirikta Barna to form words. When a Swara Barna is used with the alphabets of Byanjana Barna and Atirikta Barna, the former is used as a symbol with latter to form words. These symbols are known as Matras [2]. When a Byanjana Barna alphabet is used as a symbol with the other alphabets of Byanjana Barna, these are called Juktakhyara [2].

There are different types of software available for typing Odia language in a computer. Akruti and Microsoft Indic Language Input tool for Odia are some popularly used typing software.

This paper has concentrated on the recognition of Swara Barna, Byanjana Barna and Atirikta Barna. The alphabet is written in a document in a particular font family ***'AkrutiOriAshok-99'*** in a particular font size in bold style. The font sizes that are considered in this paper are 18, 20, 22, 24, 26, 28, 36, 48 and 72. This document is either scanned or converted to image by software. The approach described in this paper first creates a dictionary of images written in ***'AkrutiOriAshok-99'*** font family, with font sizes 18, 20, 22, 24, 26, 28, 36, 48 and 72 and in bold style. The features of these images are extracted using Preprocessing, FindPath, GettingFeaturesLeft or GettingFeaturesRight, VisitSubQuad, RemainingSubQuad modules and the extracted features are written to the excel file using WriteToExcel module. A common feature is extracted from the extracted features from the images in dictionary using CommonFeature module. For finding a correct match for the input image in the dictionary of features, FindingMatch has been used. For finding a correct match, CheckCommonFeature module of Recognition has been used. If a correct match has not been found by the CheckCommonFeature module, then MatchCommonFeature module has been used to find a correct match. If in some cases, these two modules of Recognition are unable to find a correct match, the TraceAnotherDirection module of Recognition has been used.

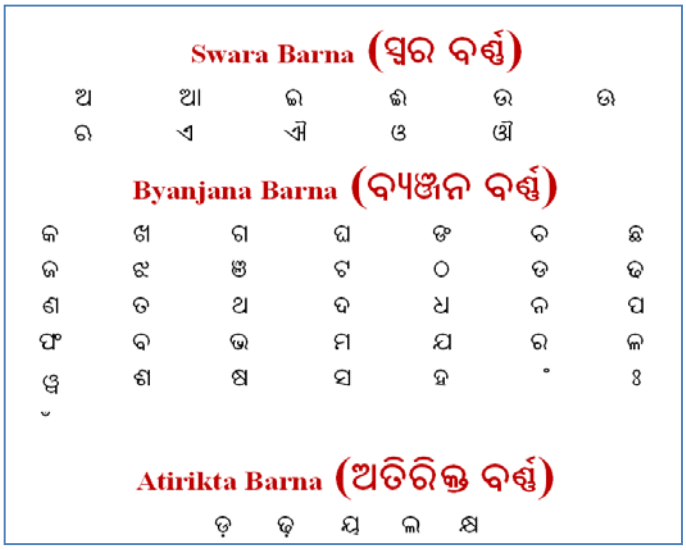


Fig. 1. Odia Alphabets.

The Preprocessing module in ***'DictionaryBuilding'*** and ***'FindingMatch'*** converts the image into gray image and then the white spaces surrounding the Odia alphabet in the gray image are removed using the Phase – 1 of RemoveNoise module of [3] (RemoveBoundarySpaces). For converting image to gray, OpenCv package of python has been used. After the elimination of white spaces from gray image, it is resized into 64 x 64 and the resultant resized image is converted to binary by using OSTU's method of thresholding [4, 5, 6]. Gray image is a type of image where intensity is stored as an 8-bit integer, hence each pixel can have intensity value ranging from 0 – 255 [7]. Binary Image is a type of image where image data is represented in terms of 0 and 1[7, 8, 9]. The basic idea for extracting features followed in this paper for DictionaryBuilding and FindingMatch is dividing the image into four quadrants and then tracing continuous path of black pixel in a particular direction in each quadrant. Experimentally, a specific direction of tracing has been agreed upon for each quadrant. The inputs to DictionaryBuilding and FindingMatch are a directory named as ***'Dictionary'*** (consists of all alphabets of Swara Barna, Byanjana Barna and Atirikta Barna of Odia language) and a directory named as ***'Input'*** (consisting of an image of Odia alphabet) respectively. The files present in ***'Dictionary'*** are accessed using os package of python [10]. The extracted features for DictionaryBuilding and FindingMatch are written in excel files, ***DictionaryFeatures.xlsx'*** and ***'InputFile.xlsx'*** respectively by using openpyxl package of python [11]. The common feature is extracted from the extracted features present in ***DictionaryFeatures.xlsx'*** by using ***Longest Common Subsequence (LCS)*** [12, 13, 14, 15, 16] and the common feature is written to the excel file, ***'CommonFeature.xlsx'*** by using openpyxl package of python. Both in DictionaryBuilding and FindingPath, Numpy package of python [17, 18] has been used for rounding off values and Matplotlib package of python [19] for sub-plotting four quadrants of the given image in one single figure (Fig. 3). Data structures like List and Dictionaries of python are used for holding multiple values. List is a data structure which behaves as a dynamic array in python and multiple values can be appended to it [20, 21]. Dictionaries consist of key values and for each key value there will be a specific value [20, 21]. The key values and values for each key value in dictionaries can be a number and can also be a string.

In other words, the proposed system concentrates on dictionary building by extracting features from the images present in ***'Dictionary'*** directory and storing the extracted features in an excel file ***'DictionaryFeatures.xlsx'***. As per the research, the same character in different font sizes results in a number of features. Therefore, it is needed to find out a common feature among all the font sizes. To achieve this ***Longest Common Subsequence (LCS)*** has been used so that there will be one common feature for a particular character. This common feature for the particular character has been stored in an excel file, ***'CommonFeature.xlsx'***. The above process is done by using phases of ***'DictionaryBuilding'***. Then feature is extracted from an input image and this feature is searched in 'CommonFeature.xlsx' by following the phases of ***'FindingMatch'*** to get a correct match. The proposed approach will help to recognise Odia characters from a scanned image or a document converted to image and these recognised characters can be written into a document and further editing can be done.

II. Related Work

The system introduced in [22], segments handwritten text into lines, from lines, words were segmented and from words characters were segmented. This system had used the water reservoir principle introduced in [23]. The input to the system was a document which was handwritten in Odia. To segment lines, the document was divided to find vertical stripes. Based on vertical projection profile and structural features of Odia characters, text lines were segmented into words. For character segmentation, at first, characters that were connected were detected. Using water-reservoir-concept touching characters of the word were then segmented. The word segmentation module was tested on 3700 words and it was noticed that the word segmentation module had an accuracy of 98.2%. The proposed technique for the isolated and connected character identification had an average accuracy of 96.7%. From the experiment it had been noticed that, in 98.6% cases, isolated characters fall into isolated group. From the experiment it had been noticed that 96.7% accuracy was obtained from two-character touching components. The accuracy of the proposed scheme on three character touching components was 95.1%.

The system introduced in [23] uses a technique for automatic segmentation of handwritten connected numerals. This system had worked on the images of French bank checks from French Company (Itesoft). Initially, the images were in gray scale (256 levels) and they had used histogram based thresholding approach to convert the gray image into binary image. Features were extracted by using the technique called water reservoir. Reservoir was obtained by the accumulation of water poured from the top or from the bottom of the numerals. Top reservoirs were formed when water was poured from the top and bottom reservoirs were formed when water was poured from the top after rotating the component by 180^{o}. Water reservoirs were the white regions of the component.

The features that were considered in the scheme were: number of reservoirs, position of reservoirs with respect to bounding box of the touching pattern, shape and size of the reservoirs, centre of gravity of the reservoirs and relative positions of the reservoirs. The segmentation result was verified manually and observed that 94.8% of the connected numerals were accurately segmented.

The system described in [24] recognises odia compound character by analysing strokes. The approach had identified 12 strokes that are enough to describe any Odia character. The input character was resized into a 60 x 60 image and then divided into nine equal halves called zones. Each zone consists of some strokes. There are nine zones and 12 strokes so; each feature vector of the character was represented in a 1 x 108. The value of similarity between strokes and zone were arranged in a vector format. Structural Similarity Index had been used as it is based on the concept that the structure of the image is independent of the illumination. The training set had been prepared from the 211 classes of Kalinga font. The system was implemented in windows machine and on MATLAB platform. The independent character recognition accuracy was achieved as 92%. The system also covers many test samples of degraded Kalinga characters. A complete OCR was also designed to work on scanned text document.

The approach described in [25] deals with handwritten Odia character recognition. This system has two level of classification. The input to the first level of classification was a cropped image. Then the input image was binarized followed by thinning. The mid value of the image was found. Then the image was divided into three equal halves row wise and two halves column wise, making it six zones. The distance between the pixel value and the centroid was calculated and this was done for all pixels for a zone and then average distance was calculated for that zone. The angle between image centroid and the pixel was calculated and this was done for each pixel in a zone. Then the average of the angles was calculated. In second-level classification, the cropped image was taken as input and it was divided into nine zones. Then the same procedure that was carried out in first-level classification was also followed in second-level classification. The first-level classification output six average distances and six average angles. The second-level classification also output nine standard deviations, nine average distances and nine average angles. Then Artificial Neural network was used for classification.

The system introduced in [26] considered each character as composition of sequence of high-level strokes and low-level strokes. They had identified low-level strokes in the system explained in [27]. In [26], they had identified forty eight visually non-redundant high-level strokes which form the maximum of a Gujarati character. Each high-level stroke is a combination of point, curves and lines. The proposed method start scanning from the center region of the character in left to right order and extract all junction points. The 3 x 3 neighbourhood of each junction point was then scanned in clockwise order to obtain the starting point of each high-level stroke. The high-level stroke ends at endpoint or until next junction point is not reached using contour tracing method. The system had used finite state machine to identify high-level stroke. For classification, the system had used Naive Bayes Classifier and Hidden Markov Model. The overall accuracy achieved using Naive Bayes Classifier and Hidden Markov Model was 93.26% and 96.87% respectively.

III. System Architecture

This approach consists of ***'DictionaryBuilding'*** and ***'FindingMatch'*** parts. The output of the above two parts are given as input to the Recognition module to find a correct match. The overall system architecture has been shown in Fig. 2

A. DictionaryBuilding

This part deals with building dictionary of features extracted from the dictionary of images of Odia alphabets which are created by scanning a document or a document converted to image by using software, both written in a font family, ***'AkrutiOriAshok-99'*** in a particular font size. The different font sizes used are 18, 20, 22, 24, 26, 28, 32, 48 and 72. For a particular font size, images of Odia alphabets of that font size are stored in a directory. Hence, nine directories are created as nine different font sizes have been used. These nine directories are stored in a directory named as ***'Dictionary'***.

The input to the ***'DictionaryBuilding'*** is the ***'Dictionary'*** directory. The directories in ***'Dictionary'*** and the image files in each directory are accessed using os package of Python. Each image file goes through ***'Feature Extraction in DictionaryBuilding'***.

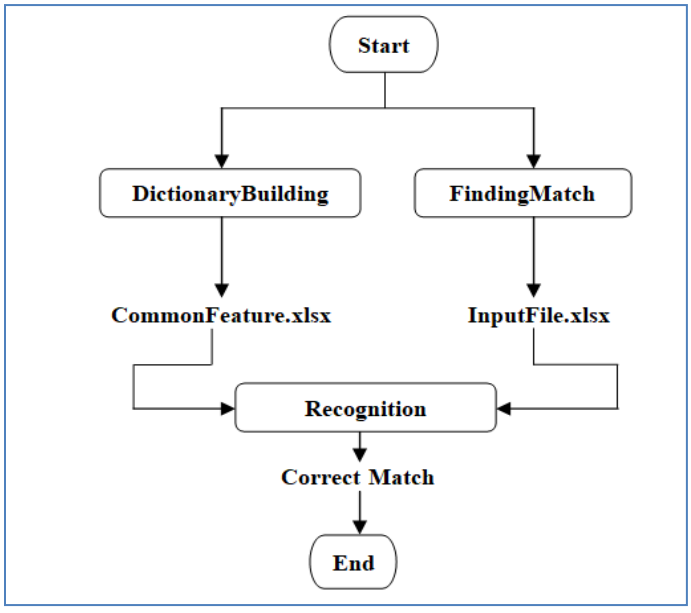


Fig. 2. System Architecture.

1) Feature Extraction in DictionaryBuilding

The ***'Dictionary'*** directory consists of nine directories, each directory dedicated to a particular font size. For example, the directory dedicated to font size 18 consists of images of each Odia alphabet written in font size 18, directory dedicated to font size 20 consist of images of each Odia alphabet written in font size 20 and so on. All the images in all these directories of ***'Dictionary'*** undergoes Preprocessing, FindPath,

GettingFeaturesRight or GettingFeaturesLeft, RemaingSubQuad, VisitSubQuad modules to extract the features of the images and these extracted features are written into an excel file using WriteToExcel Module. For each directory in the ***'Dictionary'***, a sheet is created in the excel file named as ***'DictionaryFeatures.xlsx'*** and the features are written in that sheet. The overall process of feature extraction of ***'Dictionary'*** images has been shown in Fig. 5.

a) Preprocessing Module

The input to the Preprocessing module is the directory ***'Dictionary'.***

Algorithm:

Input: Directory ***'Dictionary'***

For each image in the directories of the ***'Dictionary',*** the following steps have been followed:

1. The image is converted to gray image.
2. The white spaces that surround the text in the gray image are removed using Phase – I of RemoveNoise module of [3], that is, RemoveBoundarySpaces. This gives an image that consists of Odia alphabet only.
3. After white spaces have been removed, the image is resized into 64 x 64 by using inter-cubic interpolation.
4. The resized image consisting of Odia alphabet only is then converted into a binary image named as ***'BinaryImage'*** using OSTU's method. In ***'BinaryImage'***, the pixels that form the Odia alphabet are called black pixels and they are represented as 0 whereas the pixels that form the other areas of the ***'BinaryImage'*** are called white pixels and they are represented as 1.
5. The ***'BinaryImage'*** is divided into two equal parts, both horizontally and vertically. In this way, this image is divided into four equal quadrants. The dimension of this image is m x n (m = 64 and n = 64), where ***'m'*** is the number of rows and ***'n'*** is the number of columns. The row that equally divides the ***'BinaryImage'*** horizontally is named as ***'MidRow'*** and it is found out by using the following formula:

$$MidRow = \lceil m/2 \rceil$$

The column that equally divides the ***'BinaryImage'*** vertically is named as ***'MidCol'*** and it is found out by using the following formula:

$$MidCol = \lceil n/2 \rceil$$

6. The four quadrants are found out from the ***'BinaryImage'*** by using ***'MidRow'*** *and* ***'MidCol'***. The four quadrants 1st, 2nd, 3rd and 4th are named as B, C, D and E respectively. The four quadrants are shown in Fig. 3.

B = BinaryImage[0 : MidRow-1, 0 : MidCol-1]
C = BinaryImage[0 : MidRow-1, MidCol : n]
D = BinaryImage[MidRow : m, 0 : MidCol-1]
E = BinaryImage[MidRow : m, MidCol : n]

7. Call ***FindPath(quadNo, quadrant, DicItem, DicInnerItem, DataPath, shortName)*** for the quadrants B, C, D and E where,

quadNo is the number of quadrant among the four quadrants. Here, quadNo = 1, 2, 3, 4
quadrant can be B, C, D and E
DicItem is the ***dth*** directory in the directory ***'Dictionary'***. DicItem = 1, 2, 3, 4, 5, 6, 7, 8 and 9. For each value of DicItem, a sheet in the excel file named as ***'DictionaryFeatures.xlsx'*** is created named with the value of DicItem. For example, if DicItem = 1 then a sheet named '1' is created in the excel file.
DicInnerItem is the ***ith*** item of the ***DicItemth*** directory of ***'Dictionary'***. Each ***'DicInnerItem'*** is an image file. DicInnerItem = 1, 2, 3,....., num where ***'num'*** is the total number of image files in the ***DicItemth*** directory.
DataPath is the absolute path of the excel file, ***'DictionaryFeatures.xlsx',*** where the features are being written.
shortName is the name of the image file present in any ***DicItemth*** directory of ***'Dictionary'***.

Suppose quadNo = 1, quadrant = B, DicItem = 2, DicInnerItem = 12 then a sheet named ***'2'*** will be created in the excel file, ***'DictionaryFeatures.xlsx'*** whose path has been provided in the ***'DataPath'*** parameter, and then the extracted feature is being written in the '***12th'*** row (as DicInnerItem = 12) and ***'1st'*** column (as quadNo = 1) of the sheet. The value in the parameter ***'shortName'*** is written in the fifth column.

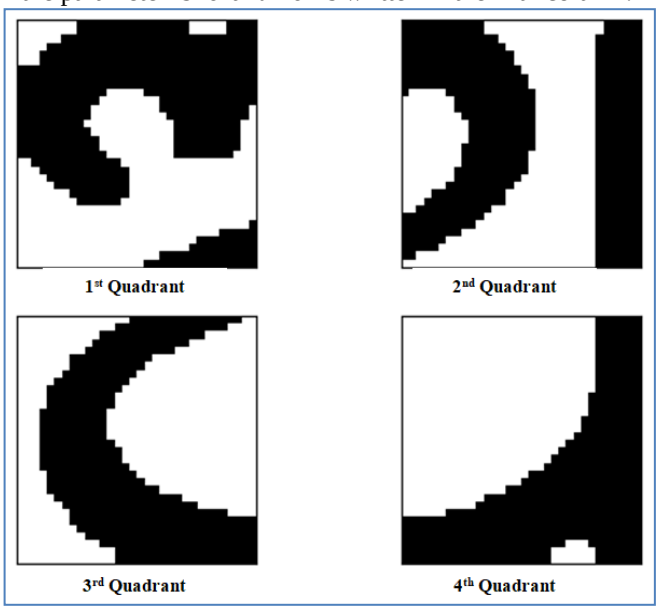


Fig. 3. An Odia Alphabet Divided into Four Quadrants.

b) FindPath Module

The steps of FindPath Module are performed for each of the quadrants B, C, D and E. The idea of this module is that the scanning of each quadrant is started from a particular corner and also scanned in a particular direction to extract the features. The scanning of Quadrant B (quadNo = 1) is started from the leftmost and bottom-most corner and when the first black pixel is found, the 'I' and 'J' (co-ordinates of the first black pixel) values are passed to GettingFeaturesRight module for scanning towards right. The scanning of Quadrant C (quadNo = 2) is started from the topmost and leftmost corner

and it is scanned towards right using GettingFeaturesRight module. The scanning of Quadrant D (quadNo = 3) is started from the bottom-most and right-most corner and it is scanned towards left using GettingFeaturesLeft module. The scanning of Quadrant E (quadNo = 4) is started from the bottom-most and left-most corner and it is scanned towards right using GettingFeaturesRight module.

row = Number of rows of quadrant

col = Number of columns of quadrant

Algorithm:

FindPath(quadNo, quadrant, DicItem, DicInnerItem, DataPath, shortName)

1. SET I = row – 1
2. SET J = 0
3. IF quadNo = 1 THEN GO TO STEP 4
4. REPEAT STEP 5 WHILE J < col
5. REPEAT STEP 6 WHILE I > 0
6. IF quadrant[I][J] = 0 THEN GO TO STEP 7
7. CALL GettingFeaturesRight(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
8. SET I = 0
9. SET J = 0
10. IF quadNo = 2 THEN GO TO STEP 11
11. REPEAT STEP 12 WHILE I < row
12. REPEAT STEP 13 WHILE J < col
13. IF quadrant[I][J] = 0 THEN GO TO STEP 14
14. CALL GettingFeaturesRight(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
15. SET I = row – 1
16. SET J = col – 1
17. IF quadNo = 3 THEN GO TO STEP 18
18. REPEAT STEP 19 WHILE I > 0
19. REPEAT STEP 20 WHILE J > 0
20. IF quadrant[I][J] = 0 THEN GO TO STEP 21
21. CALL GettingFeaturesLeft(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
22. SET I = row – 1
23. SET J = 0
24. IF quadNo = 4 THEN GO TO STEP 25
25. REPEAT STEP 26 WHILE I > 0
26. REPEAT STEP 27 WHILE J < col
27. IF quadrant[I][J] = 0 THEN GO TO STEP 28
28. CALL GettingFeaturesRight(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
29. EXIT

c) GettingFeaturesLeft Module

When the first black pixel is found in FindPath Module while scanning the quadrant from the specified corner, the coordinates of the pixel (I value and J value) are passed to this module to get a continuous trace of black pixels in the specified quadrant. This module scans the quadrant towards the left starting from the first black pixel. This module is used in quadrant D.

'quadrant', ***'quadNo'***, ***'DicItem'***, ***'DicInnerItem'***, ***'DataPath'*** and ***'shortName'*** are explained in the step 7 of preprocessing module.

'LSubQuad' will contain the final feature extracted from a particular quadrant.

'I' and ***'J'*** consists of the row and column number of the first black pixel obtained in a particular quadrant using FindPath module.

row1 = Number of rows of quadrant

col1 = Number of columns of quadrant

Algorithm:

GettingFeaturesLeft(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)

1. IF J = 0 THEN DO STEPS FROM 2 TO 5
2. LSubQuad = CALL VisitSubQuad(I, J, 0, row1, 0, col1)
3. CALL RemainingSubQuad(quadrant)
4. CALL WriteToExcel(DicItem, DicInnerItem, quadNo, LSubQuad, DataPath, shortName)
5. RETURN
6. ELSE DO STEPS 7 OR 12 OR 17 OR 22, WHICHEVER SATISFIES CONDITION FIRST
7. IF quadrant[I – 1, J - 1] = 0 THEN DO STEPS FROM 8 TO 11
8. I = I – 1
9. J = J – 1
10. LSubQuad = CALL VisitSubQuad(I, J, 0, row1, 0, col1)
11. CALL GettingFeaturesLeft(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
12. ELSE IF quadrant[I, J - 1] = 0 THEN DO STEPS FROM 13 TO 16
13. I = I
14. J = J – 1
15. LSubQuad = CALL VisitSubQuad(I, J, 0, row1, 0, col1)
16. CALL GettingFeaturesLeft (I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
17. ELSE IF quadrant[I + 1, J - 1] = 0 THEN DO STEPS FROM 18 TO 21
18. I = I + 1
19. J = J – 1
20. LSubQuad = CALL VisitSubQuad(I, J, 0, row1, 0, col1)
21. CALL GettingFeaturesLeft (I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
22. ELSE IF (quadrant[I – 1, J - 1] = 1 AND quadrant[I, J - 1] = 1 AND quadrant[I + 1, J - 1] = 1) OR (J = 0) OR (I = 0) OR (I = row1 – 1) OR (J = col1 – 1) THEN DO STEPS FROM 23 TO 25

23. CALL RemainingSubQuad(quadrant)
24. CALL WriteToExcel(DicItem, DicInnerItem, quadNo, LSubQuad, DataPath, shortName)
25. RETURN
26. EXIT

d) GettingFeaturesRight Module

The aim of this module is to get a continuous trace of black pixels scanning from left to right. When the first black pixel is found in FindPath Module while scanning the quadrant from the specific corner, the coordinates of the pixel (I value and J value) are passed to this module to get a continuous trace of black pixels in the specified quadrant. This module is used in quadrants B, C and E.

'quadrant', ***'quadNo'***, ***'DicItem'***, ***'DicInnerItem'***, ***'DataPath'*** and ***'shortName'*** are explained in the step 7 of Preprocessing module.

'LSubQuad' will contain the final feature extracted from a particular quadrant.

'I' and ***'J'*** consists of the row and column number of the first black pixel obtained in a particular quadrant using FindPath module.

row1 = Number of rows of quadrant

col1 = Number of columns of quadrant

Algorithm:

GettingFeaturesRight(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
1. IF J = col1 - 1 THEN DO STEPS FROM 2 TO 5
2. LSubQuad = CALL VisitSubQuad(I, J, 0, row1, 0, col1)
3. CALL RemainingSubQuad(quadrant)
4. CALL WriteToExcel(DicItem, DicInnerItem, quadNo, LSubQuad, DataPath, shortName)
5. RETURN
6. ELSE DO STEPS 7 OR 12 OR 17 OR 22 WHICHEVER CONDITION SATISFIES FIRST
7. IF quadrant[I – 1, J + 1] = 0 THEN DO STEPS FROM 8 TO 11
8. I = I – 1
9. J = J + 1
10. LSubQuad = CALL VisitSubQuad(I, J, 0, row1, 0, col1)
11. CALL GettingFeaturesRight(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
12. ELSE IF quadrant[I, J + 1] = 0 THEN DO STEPS FROM 13 TO 16
13. I = I
14. J = J + 1
15. LSubQuad = CALL VisitSubQuad(I, J, 0, row1, 0, col1)
16. CALL GettingFeaturesRight(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
17. ELSE IF quadrant[I + 1, J + 1] = 0 THEN DO STEPS FROM 18 TO 21
18. I = I + 1
19. J = J + 1
20. LSubQuad = CALL VisitSubQuad(I, J, 0, row1, 0, col1)
21. CALL GettingFeaturesRight(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
22. ELSE IF (quadrant[I – 1, J + 1] = 1 AND quadrant[I, J + 1] = 1 AND quadrant[I + 1, J + 1] = 1) OR (J = 0) OR (I = 0) OR (I = row1 – 1) OR (J = col1 – 1) THEN DO STEPS FROM 23 TO 25
23. CALL RemainingSubQuad(quadrant)
24. CALL WriteToExcel(DicItem, DicInnerItem, quadNo, LSubQuad, DataPath, shortName)
25. RETURN
26. EXIT

e) VisitSubQuad Module

Each of the four quadrants **B, C, D** and **E** is again divided into four sub-quadrants named as a, b, c and d. This module uses a list data structure named as ***'subQuad'*** and appends name of the sub-quadrant (***'a'*** or ***'b'*** or ***'c'*** or ***'d'***) for each black pixel found while scanning by GettingFeaturesLeft or GettingFeaturesRight module including the first black pixel found in the FindPath Module in ***'SubQuad'***. The sub-quadrants of each quadrant have been shown in Fig. 4. The value returned by this module is stored in ***'LSubQuad'*** of either GettingFeaturesLeft or GettingFeaturesRight module. The final value in ***'LsubQuad'*** is the feature extracted for a particular quadrant (***'B'*** or ***'C'*** or ***'D'*** or ***'E'***).

'srow' is the row number where the quadrant starts.

'erow' is the row number where the quadrant ends.

'scol' is the column number where the quadrant starts.

'ecol' is the column number where the quadrant ends.

$$frow = \left[\left((erow - srow)\right)/2\right]$$

$$fcol = \left[\left((ecol - scol)\right)/2\right]$$

Algorithm:

VisitSubQuad(I, J, srow, erow, scol, ecol)
1. IF (I >= srow AND I <= frow – 1) AND (J >= scol AND J <= fcol – 1) GO TO STEP 2
2. APPEND ***'a'*** in ***'subQuad'***
3. ELSE IF (I >= srow AND I <= frow – 1) AND (J >= fcol AND J <= ecol) GO TO STEP 4
4. APPEND ***'b'*** in ***'subQuad'***
5. ELSE IF (I >= frow AND I <= erow) AND (J >= fcol AND J <= ecol) GO TO STEP 6
6. APPEND ***'c'*** in ***'subQuad'***
7. ELSE IF (I >= frow AND I <= erow) AND (J >= scol AND J <= fcol - 1) GO TO STEP 8
8. APPEND ***'d'*** in ***'subQuad'***
9. RETURN ***'subQuad'***
10. EXIT

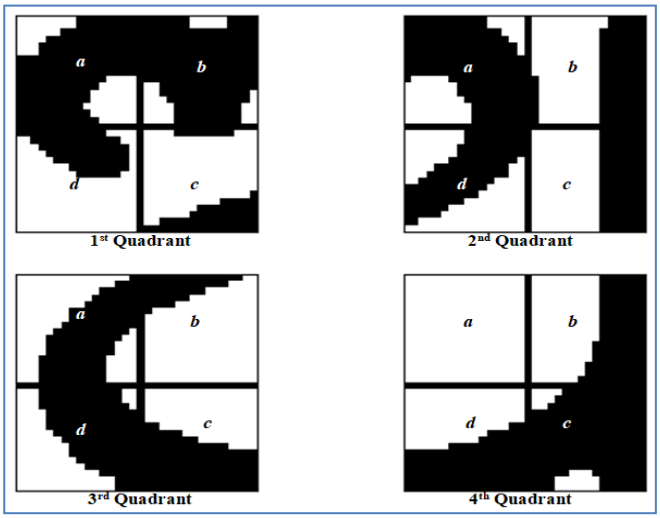


Fig. 4. Sub-Quadrants of Four Quadrants.

f) RemainingSubQuad Module

A continuous trace of black pixels is found by either GettingFeaturesLeft or GettingFeaturesRight starting from the first black pixel found in FindPath module and the name of the sub-quadrant is appended in ***'LSubQuad'***. But the continuous trace of black pixels may not have accessed some part of the quadrant (***'B'*** or ***'C'*** or ***'D'*** or ***'E'***). To ensure, all parts of the quadrant have been accessed, the remaining parts are accessed using RemainingSubQuad Module. First, this module checks if all the sub-quadrants have been accessed for the black pixels. This is done by checking the contents of the ***'LSubQuad'*** list. In other words, if a sub-quadrant does not have any black pixel then that sub-quadrant is not allowed to be present in the ***'LSubQuad'*** list. If the name of the all sub-quadrants that have black pixels have appeared at least once in ***'LSubQuad'*** then, RemainingSubQuad exits, otherwise, RemainingSubQuad is called recursively to scan the sub-quadrants until all sub-quadrants that have black pixels have been scanned and stored in the ***'LSubQuad'***.

'quadrant' consists of ***'B'***, ***'C'***, ***'D'*** and ***'E'***.

Algorithm:

RemainingSubQuad(quadrant)

1. IF ***'a'***, ***'b'***, ***'c'*** and ***'d'*** all are in ***'LSubQuad'*** THEN
2. GO TO STEP 20
3. ELSE
4. IF ***'a'*** IS NOT IN ***'LSubQuad'*** THEN DO STEP 5 TO 7
5. SCAN black pixels of quadrant from top-most and left-most corner to find the first black pixel and from there scan towards right following the similar procedure as in GettingFeaturesRight to find the continuous trace of black pixel.
6. For each black pixel in the continuous trace APPEND ***'a'*** in ***'LSubQuad'.***
7. CALL RemainingSubQuad(quadrant)
8. ELSE IF ***'b'*** IS NOT IN ***'LSubQuad'*** THEN DO STEP 9 TO 11
9. SCAN black pixels of quadrant from top-most and right-most corner to find the first black pixel and from there scan towards left following the similar procedure as in GettingFeaturesLeft to find the continuous trace of black pixel.
10. For each black pixel in the continuous trace APPEND ***'b'*** in ***'LSubQuad'.***
11. CALL RemainingSubQuad(quadrant)
12. ELSE IF ***'c'*** IS NOT IN ***'LSubQuad'*** THEN DO STEP 13 TO 15
13. SCAN black pixels of quadrant from bottom-most and right-most corner to find the first black pixel and from there scan towards left following the similar procedure as in GettingFeaturesLeft to find the continuous trace of black pixel.
14. For each black pixel in the continuous trace APPEND ***'c'*** in ***'LSubQuad'.***
15. CALL RemainingSubQuad(quadrant)
16. ELSE IF ***'d'*** IS NOT IN ***'LSubQuad'*** THEN DO STEP 17 TO 19
17. SCAN black pixels of quadrant from bottom-most and left-most corner to find the first black pixel and from there scan towards right following the similar procedure as in GettingFeaturesRight to find the continuous trace of black pixel.
18. For each black pixel in the continuous trace APPEND ***'d'*** in ***'LSubQuad'.***
19. CALL RemainingSubQuad(quadrant)
20. EXIT

g) WriteToExcel Module

When a quadrant (***'B'*** or ***'C'*** or ***'D'*** or ***'E'***) is scanned completely for tracing black pixels, all the features of the quadrant are extracted in the form of a, b, c and d and stored in ***'LSubQuad'***. The contents of the ***'LSubQuad'*** are concatenated to present the features in a string format. This string value is written in the ***'DicItemth'*** sheet, ***'DicInnerItemth'*** row and ***'quadNoth'*** column of the excel file, ***'DictionaryFeatures.xlsx'***. The path of the excel file is stored in ***'DataPath'*** parameter. The ***'shortName'*** is the file name and it is written in the fifth column and ***'DicInnerItemth'*** row of the ***'DicItemth'*** sheet of the excel file. All the extracted features are written in the excel file by using the openpyxl package of Python. This excel file contains the features of each alphabet in all font sizes (18, 20, 22, 24, 26, 28, 32, 48 and 72). For example, the alphabet ಅ in font size 18, 20, 22, 24, 26, 28, 32, 48 and 72 is written in the first row of sheet named 1, 2, 3, 4, 5, 6, 7, 8 and 9 respectively.

Algorithm:

WriteToExcel(DicItem, DicInnerItem, quadNo, LSubQuad, DataPath, shortName)

1. INITIALIZE ***'S'*** to an empty string
2. A = 0
3. REPEAT STEP 4 WHILE A < LENGTH(LSubQuad)
4. S = S + LSubQuad[A]
5. For each value of DicItem, CREATE a new sheet named with the value of the DicItem.

6. IF the current value of DicItem is same as the value in the previous iteration THEN a new sheet is not created and the feature in string format is written in the current sheet of the excel file, ***'DictionaryFeatures.xlsx'.***
7. ELSE CREATE a new sheet for writing extracted features of the alphabets of next font size in the excel file, ***'DictionaryFeatures.xlsx'***.
8. EXIT

h) CommonFeature Module

When all the features have been extracted and written in the excel file, ***'DictionaryFeatures.xlsx'*** for all alphabets in all font sizes using WriteToExcel module then, a common feature is found from all the extracted features of an alphabet in different font sizes. For this, CommonFeature module is used. This module finds the ***LCS (Longest Common Subsequence)*** of all the features of a particular alphabet in different font sizes to find the common feature. LCS is a way of finding longest common sub-sequences from a set of sequences.The common feature found using LCS is written in another excel file named as ***'CommonFeature.xlsx'*** which consists of only one sheet.

'fiQuList', ***'SeQuList'***, ***'ThQuList'*** and ***'FoQuList'*** are the lists that contains features of first quadrant (B), second quadrant (C), third quadrant (D) and fourth quadrant (E) of a particular alphabet in different font sizes respectively.

'row' is the total number of rows present in the excel file, ***'DictionaryFeatures.xlsx'*** which contains features of all alphabets in a sheet. The value of ***'row'*** is same in all sheets of the excel file, ***'DictionaryFeatures.xlsx'***.

'sheet' is the total number of sheets present in the excel file, ***'DictionaryFeatures.xlsx'*** which contains features of all alphabets in different font sizes (In this research, sheet = 9 as nine different font sizes are considered).

'Text1', ***'Text2'***, ***'Text3'*** and ***'Text4'*** are strings.

'sr' is initialized to 0.

'sh' is initialized to 0.

Algorithm:

CommonFeature()
1. REPEAT STEP 2 WHILE sr < row
2. REPEAT STEPS FROM 3 TO 6 WHILE sh < sheet
3. APPEND the feature in first column of ***'sr^{th}'*** row of ***'sh^{th}'*** sheet of ***'DictionaryFeatures.xlsx'*** in ***'fiQuList'.***
4. APPEND the feature in second column of ***'sr^{th}'*** row of ***'sh^{th}'*** sheet of ***'DictionaryFeatures.xlsx'*** in ***'SeQuList'.***
5. APPEND the feature in third column of ***'sr^{th}'*** row of ***'sh^{th}'*** sheet of ***'DictionaryFeatures.xlsx'*** in ***'ThQuList'.***
6. APPEND the feature in fourth column of ***'sr^{th}'*** row of ***'sh^{th}'*** sheet of ***'DictionaryFeatures.xlsx'*** in ***'FoQuList'.***
7. Text1 = fiQuList[0]
8. f = 0
9. REPEAT STEP 10 WHILE f < (LENGTH (fiQuList) – 1)
10. Text1 = FindLCS(Text1, fiQuList[f + 1]
11. Text2 = SeQuList[0]
12. f = 0
13. REPEAT STEP 14 WHILE f < (LENGTH (SeQuList) – 1)
14. Text2 = FindLCS(Text2, SeQuList[f + 1]
15. Text3 = ThQuList[0]
16. f = 0
17. REPEAT STEP 18 WHILE f < (LENGTH (ThQuList) – 1)
18. Text3 = FindLCS (Text3, ThQuList[f + 1])
19. Text4 = FoQuList[0]
20. f = 0
21. REPEAT STEP 22 WHILE f < (LENGTH (FoQuList) – 1)
22. Text4 = FindLCS (Text4, FoQuList[f + 1])
23. WRITE the value of ***'Text1'*** in the first column of ***'sr^{th}'*** row of the excel file, ***'CommonFeature.xlsx'***.
24. WRITE the value of ***'Text2'*** in the second column of ***'sr^{th}'*** row of the excel file, ***'CommonFeature.xlsx'***.
25. WRITE the value of ***'Text3'*** in the third column of ***'sr^{th}'*** row of the excel file, ***'CommonFeature.xlsx'***.
26. WRITE the value of ***'Text4'*** in the fourth column of ***'sr^{th}'*** row of the excel file, ***'CommonFeature.xlsx'***.
27. Clear all the lists ***'FiQuList'***, ***'SeQuList'***, ***'ThQuList'*** and ***'FoQuList'.***
28. INITIALIZE ***'Text1'***, ***'Text2'***, ***'Text3'*** and ***'Text4'*** to empty string.
29. EXIT

i) Longest Common Subsequence

This module finds and returns the LCS (Longest Common Subsequence) of ***'String1'*** and ***'String2'***. The LCS algorithm has been implemented using Dynamic Programming in this paper. If ***'String1'*** and ***'String2'*** are equal then ***'String1'*** is stored in ***'ls'*** and it is returned, otherwise the LCS of ***'String1'*** and ***'String2'*** is found out and it is stored in ***'revLs'*** and returned. The two arrays ***'LcsForm'*** and ***'b'*** are used to store the length of the LCS and the traversing direction of the LCS respectively in each column of each row. The ***'s'***, ***'u'*** and ***'l'*** denote ***'towards diagonal'***, ***'towards upper'*** and ***'towards left'*** directions respectively. After all the values of ***'LcsForm'*** and ***'b'*** are found out, both arrays are scanned from the bottom-most corner and right-most side to get the value of **I** and **J** where ***LcsForm[I][J] = MaxValue and b[I][J] = 's'*** and for each ***'s'*** in ***'b'*** array, the common item in both the strings (String1 and String2) is appended in ***'ls'***. The ***'MaxValue'*** is the maximum length of LCS in '***LcsForm'***. At last ***'ls'*** is reversed and the result is stored in ***'revLs'***.

Algorithm:

FindLCS(String1, String2)
1. IF String1 = String2 THEN DO STEP 2 TO 3
2. APPEND ***'String1'*** in the list named as ***'ls'***.
3. RETURN ***'ls'***.
4. ELSE DO FROM STEP 5 TO 37
5. m = LENGTH(String1)

6. n = LENGTH(String2)
7. INITIALIZE the array ***'LcsForm'*** with dimensions (m + 1, n + 1) to zero.
8. INITIALIZE the array ***'b'*** with dimensions (m + 1, n + 1) to zero.
9. I = 0
10. J = 0
11. REPEAT STEP 12 WHILE I < (m + 1)
12. REPEAT STEP 13 or 16 or 19 WHICEVER SATISFIES THE CONDITION FIRST WHILE J < (n + 1)
13. IF String1[I – 1] = String2[J – 1] THEN DO STEP 14 TO 15
14. LcsForm[I][J] = LcsForm[I][J] + 1
15. b[I][J] = ***'s'***
16. ELSE IF LcsForm[I – 1][J] >= LcsForm[I][J – 1] THEN DO STEP 17 TO 18
17. LcsForm[I][J] = LcsForm[I – 1][J]
18. b[I][J] = ***'u'***
19. ELSE DO STEP 20 TO 21
20. LcsForm[I][J] = LcsForm[I][J – 1]
21. b[I][J] = ***'l'***
22. Find the maximum value in the array ***'LcsForm'*** and it is stored in ***'MaxValue'***.
23. Search the array ***'LcsForm'*** from right-most side and bottom-most corner of the array and find the value of I and J in the array where LcsForm[I][J] = ***MaxValue*** and b[I][J] = ***'s'***.
24. After values of I and J are found for LcsForm[I][J] = ***MaxValue*** and b[I][J] = ***'s'***, DO STEP 25
25. REPEAT STEP 26 or 30 or 33 WHICHEVER SATISFIES THE CONDITION FIRST WHILE I > 0 AND J > 0
26. IF b[I][J] = ***'s'*** THEN DO STEP 27 TO 29
27. APPEND the value in String1[I][J] in the list ***'ls'***.
28. I = I – 1
29. J = J – 1
30. ELSE IF b[I][J] = ***'u'*** DO STEP 31 TO 32
31. I = I – 1
32. J = J
33. ELSE IF b[I][J] = ***'l'*** DO STEP 34 TO 35
34. I = I
35. J = J – 1
36. REVERSE the items of the list ***'ls'*** and store it in ***'revLs'***.
37. RETURN ***'revLs'***.
38. EXIT

Hence, the final output of the ***'DictionaryBuilding'*** is the ***'CommonFeature.xlsx'*** excel file which consists of the common feature for each alphabet extracted from features present in ***'DictionaryFeatures.xlsx'.*** For example, the final common features for the alphabet ଅ are:

1st quadrant – aaaaaaaaaaaaabbbbbbccccccccccccc

2nd quadrant – aaaaaaabbbbbccccccddddddddddddd

3rd quadrant – ccccccccccccccddddaaaabbbbbbbbbbbbaaaaaaaaa

4th quadrant - ddddddddddddddccccbbbbbbbbbb

B. FindingMatch

This part deals with finding a correct match from the dictionary of common features stored in the excel file, ***'CommonFeature.xlsx'*** when an image of Odia alphabet is provided as input. This input image is stored in a directory named as ***'Input'***. The ***'FindingMatch'*** part undergoes through two phases: ***'Feature Extraction'*** and ***'Recognition'***.

1) Feature Extraction

This phase undergoes through seven modules for extracting features from the input image present in the directory ***'Input'*** and the features are written to an excel file named as ***'InputFile.xlsx'***. The different modules are: Preprocessing, FindPath, GettingFeaturesRight or GettingFeaturesLeft, VisitSubQuad and RemainingSubQuad for extracting features from the input image and the features are written in the excel file using WriteToExcel Module. The overall process of feature extraction of Input image has been shown in Fig. 5.

a) Preprocessing Module

The steps in this module are same as described in Preprocessing module of ***'DictionaryBuilding'*** except the values passed to the parameters in FindPath module. The input to this module is the directory ***'Input'*** consisting of an image of Odia alphabet. The input image is converted to gray image. The white spaces surrounding the Odia alphabet in the gray image are removed using the Phase – 1 of RemoveNoise module of [3] (RemoveBoundarySpaces). Then the resultant image is resized into the dimension p x q (p = 64 and q = 64) where, ***'p'*** is the number of rows and ***'q'*** is the number of columns**.** Then the resized image is converted to binary image named as ***'BinayImageIn'.*** The row that equally divides the ***'BinayImageIn'*** horizontally is named as ***'MidRow'*** and it is found out by using the following formula:

$$MidRow = \lceil {}^{p}/_{2} \rceil$$

The column that equally divides the ***'BinayImageIn'*** vertically is named as ***'MidCol'*** and it is found out by using the following formula:

$$MidCol = \lceil {}^{q}/_{2} \rceil$$

The four quadrants are found out from ***'BinayImageIn'*** in the following way:

U = BinaryImageIn[0 : MidRow-1, 0 : MidCol-1]

V = BinaryImageIn[0 : MidRow-1, MidCol : n]

W = BinaryImageIn[MidRow : m, 0 : MidCol-1]

X = BinaryImageIn[MidRow : m, MidCol : n]

'U', ***'V'***, ***'W'*** and ***'X'*** are the first, second, third and fourth quadrant respectively.

Then CALL FindPath(quadNo, quadrant, DicItem, DicInnerItem, DataPath, shortName) for the quadrants U, V, W and X.

b) FindPath Module

The steps in this module are same as described in FindPath module of ***'DictionaryBuilding'*** except that the steps of FindPath Module are performed for each of the quadrants U, V, W and X. Here ***'DicItem'*** and ***'DicInnerItem'*** are constants and are set to 1 as the ***'Input'*** folder has no sub-directories and it has only one image at any given time. The ***'DataPath'*** parameter holds the absolute path of the excel file named as ***'InputFile.xlsx'*** and in this file features of all the four quadrants of the input image are being written. The features of ***'U'***, ***'V'***, ***'W'*** and ***'X'*** are written in first, second, third and fourth column of ***'InputFile.xlsx'*** respectively.

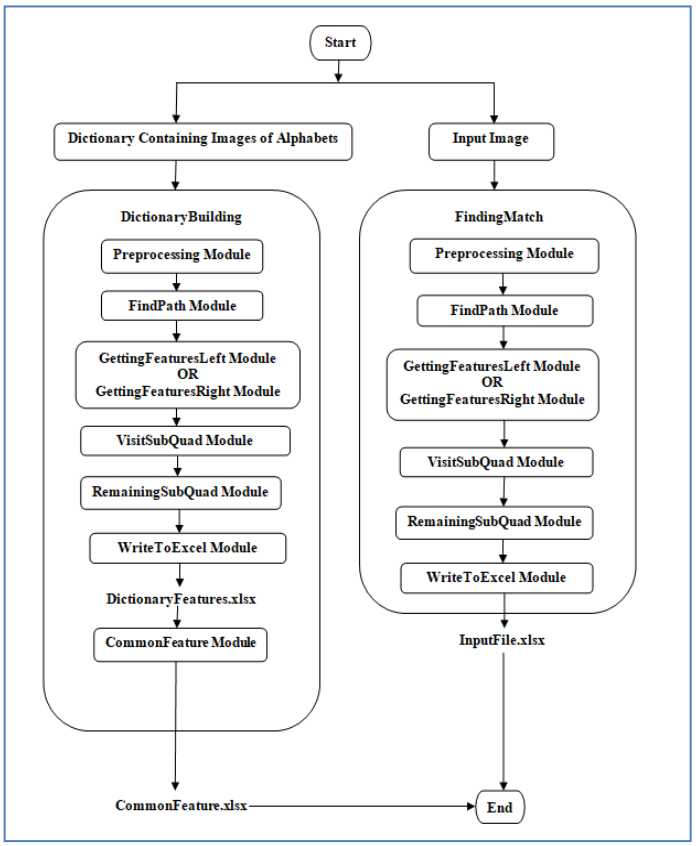


Fig. 5. Feature Extraction for Dictionary of Images and Input Image.

c) GettingFeaturesLeft Module

The steps in this module are same as described in GettingFeaturesLeft module of ***'DictionaryBuilding'*** except that the steps here are applied to W quadrant.

d) GettingFeaturesRight Module

The steps in this module are same as described in GettingFeaturesRight module of ***'DictionaryBuilding'*** except that the steps here are applied to U, V and X quadrants.

e) VisitSubQuad Module

The steps in this module are same as described in VisitSubQuad module of ***'DictionaryBuilding'*** except that the steps are applied to U, V, W and X quadrants. Similar to as explained in the VisitSubQuad module of ***'DictionaryBuilding'***, the quadrants are divided into four sub-quadrants, a, b, c and d. For each black pixel in the continuous trace, the sub-quadrant (either ***'a'*** or ***'b'*** or ***'c'*** or ***'d'***) is found out and appended in ***'subQuad'***. The value of ***'subQuad'*** is returned and set to ***'LSubQuad'*** in ***'GettingFeaturesLeft'*** or ***'GettingFeaturesRight'***, whichever has been called.

f) RemainingSubQuad Module

The steps in this module are same as described in RemainingSubQuad module of ***'DictionaryBuilding'*** except that the steps are applied to U, V, W and X quadrants. If any portions of the quadrants U, V, W and X are not covered by the continuous trace of black pixels, those remaining portions are covered by this module and the name of sub-quadrants (either ***'a'*** or ***'b'*** or ***'c'*** or ***'d'***) are appended in ***'LSubQuad'***.

g) WriteToExcel Module

The steps in this module are same as described in WriteToExcel module of ***'DictionaryBuilding'*** except that the features extracted from the quadrants U, V, W and X are written in an excel file named as ***'InputFile.xlsx'***. The absolute path of ***'InputFile.xlsx'*** is stored in the ***'DataPath'*** parameter and the file name of input image is stored in ***'shortName'***. The value in ***'shortName'*** parameter is written in the fifth column of ***'InputFile.xlsx'.*** Hence, the feature extracted from the quadrants U, V, W, X and the value in ***'shortName'*** parameter are written in the first, second, third, fourth and fifth column of the first row of the excel file, '***InputFile.xlsx'*** respectively and there is only one sheet present in the excel file as there is no sub-directories of the ***'Input'*** directory. For example, the final feature for the input image are:

1st quadrant – aaaaaaaaaaaaaaabbbbbbbbbbbbbbbbbccccccccccccccddd

2nd quadrant – aaaaaaaaaaaaabbbbbbbbcccccccddddddddddddddd

3rd quadrant – ccccccccccccccddddddaaaaaaabbbbbbbbbbbbaaaaaaaaaa

4th quadrant - ddddddddddddddccccbbbbbbbbbbb

2) Recognition

This phase undergoes through three modules: CheckCommonFeature module, MatchCommonFeature module and TraceAnotherDirection Module. The overall process of recognition has been shown in Fig. 6.

InCol: number of columns in the ***'Input.xlsx'***

ComRow: number of rows in the ***'CommonFeature.xlsx'***

'InpPat' is a list consisting of the final features of 1st, 2nd, 3rd and 4th quadrants for the ***'e^{th}'*** row of ***'Input.xlsx'.***

'QuList' is a list consisting of the final features of 1st, 2nd, 3rd and 4th quadrants for the ***'f^{th}'*** row of ***'CommonFeature.xlsx'.***

'MatchFirst' is a list consisting of the file names of the matched features obtained as the output of CheckCommonFeature module.

'MatchedRow' is a list consisting of the row numbers of the matched features in ***'CommonFeature.xlsx'.***

'MatchSecond' is a list consisting of the output of MatchCommonFeature module.

Algorithm:

Recognition( )

1. SET e = 1
2. REPEAT STEP 3 WHILE e <= InCol
3. APPEND the feature present in ***'1st'*** row and ***'eth'*** column of ***'Input.xlsx'*** in the list ***'InpPat'.***
4. SET f = 1
5. REPEAT STEPS FROM 6 TO 17 WHILE f < ComRow
6. APPEND the feature present in the ***'fth'*** row and ***'1st'*** column of ***'CommonFeature.xlsx'*** in the list ***'QuList'.***
7. APPEND the feature present in the ***'fth'*** row and ***'2nd'*** column of ***'CommonFeature.xlsx'*** in the list ***'QuList'.***
8. APPEND the feature present in the ***'fth'*** row and ***'3rd'*** column of ***'CommonFeature.xlsx'*** in the list ***'QuList'.***
9. APPEND the feature present in the ***'fth'*** row and ***'4th'*** column of ***'CommonFeature.xlsx'*** in the list ***'QuList'.***
10. ite = 1
11. REPEAT STEPS FROM 12 TO 14 WHILE ite <= LENGTH (InpPat)
12. Param1 = CALL CheckCommonFeature(InpPat[ite], QuList[ite])
13. IF Param1 = 1 THEN GO TO STEP 14
14. Param3 = Param3 + 1
15. IF Param3 = 4 THEN GO TO STEP 16
16. RETRIEVE the file name of the matched image feature present in the ***'fth'*** row and ***'5th'*** column of the ***'CommonFeature.xlsx'*** and APPEND the file name in a list named as ***'MatchFirst'*** and ***'fth'*** row number of the matched image feature in a list ***'MatchedRow'***.
17. CLEAR the list ***'QuList'***.
18. IF LENGTH(MatchFirst) > 1 THEN DO STEPS 19 TO 33
19. SET f = 1
20. REPEAT STEPS FROM 21 TO 33 WHILE f <= LENGTH (MatchFirst)
21. APPEND the feature present in the ***'(MatchedRow[f])th'*** row and ***'1st'*** column of ***'CommonFeature.xlsx'*** in the list ***'QuList'.***
22. APPEND the feature present in the ***'(MatchedRow[f])th'*** row and ***'2nd'*** column ***'CommonFeature.xlsx'*** in the list ***'QuList'.***
23. APPEND the feature present in the ***'(MatchedRow[f])th'*** row and ***'3rd'*** column ***'CommonFeature.xlsx'*** in the list ***'QuList'.***
24. APPEND the feature present in the ***'(MatchedRow[f])th'*** row and ***'4th'*** column ***'CommonFeature.xlsx'*** in the list ***'QuList'.***
25. SET ite = 1
26. REPEAT STEPS FROM 27 TO 29 WHILE ite <= LENGTH(InpPat)
27. Param2 = CALL MatchCommonFeature (InpPat[ite], QuList[ite])
28. IF Param2 = 1 THEN GO TO STEP 29
29. Param4 = Param4 + 1
30. IF Param4 = 4 THEN DO STEPS 31 TO 32
31. RETRIEVE the file name of the matched image feature present in the ***'(MatchedRow[f])th'*** row and ***'5th'*** column of the ***'CommonFeature.xlsx'*** and APPEND the file name in a list named as ***'MatchSecond'***.
32. PRINT 'MatchSecond'
33. CLEAR the list 'QuList'
34. ELSE GO TO STEP 35
35. PRINT the list 'MatchFirst'
36. EXIT

a) CheckCommonFeature Module

When feature extraction of the input image has been completed, features of all the four quadrants U, V, W and X are written in the excel file named as ***'InputFile.xlsx'***. The common feature of a particular alphabet that is extracted from all the features present in ***'DictionaryFeatures.xlsx'*** file has been written in ***'CommonFeature.xlsx'***. The common feature of first quadrant present in the first column of each row of ***'CommonFeature.xlsx'*** is searched to find whether the sequence of common feature is present in the first column of ***'InputFile.xlsx'***. The same search procedure is repeated for second, third and fourth columns of both the files, ***'CommonFeature.xlsx'*** and ***'InputFile.xlsx'.*** In other words, the second, third and fourth columns of each row of ***'CommonFeature.xlsx'*** are searched in second, third and fourth columns of ***'InputFile.xlsx'*** respectively. For this, the CheckCommonFeature(String1, String2) has been used. The presence of common features of ***'CommonFeature.xlsx'*** in the features of ***'InputFile.xlsx'*** in a continuous form or non-continuous form helps to find a correct match. If the features in the first, second, third and fourth columns of ***'InputFile.xlsx'*** are present in the first, second, third and fourth columns of ***'CommonFeature.xlsx'*** respectively in a particular row then Param1 is set to 1 otherwise it is set to 0. If Param1 = 1 then Param3 is incremented by 1. For example, if the value of Param3 is 4 after all the columns of ***'InputFile.xlsx'*** have been checked with the respective columns of ***'CommonFeature.xlsx'*** for all rows, then the file name is retrieved from the fifth column of the row that consists of matched image feature in ***'CommonFeature.xlsx'.*** The file name from fifth column of matched image feature is appended in the list ***'MatchFirst'*** and the row number of matched image feature is appended in the list ***'MatchedRow'***. In some cases, the list ***'MatchFirst'*** have more than one correct match and in these cases ***'MatchCommonFeature'*** module is called.

b) MatchCommonFeature Module

This module is used when the list ***'MatchFirst'*** (output of ***'CheckCommonFeature')*** consists of more than one match. In this module, the common features from the first, second, third and fourth columns of the each row number that is present in the list ***'MatchedRow'*** are retrieved from ***'CommonFeature.xlsx'*** and appended in the list ***'QuList'***. The features present in the first, second, third and fourth columns present in ***'InputFile.xlsx'*** are retrieved and appended in the list ***'Inpat'.*** Then the LCS (Longest Common Sequence) of the two strings, ***'Str1'*** and ***'Str2'*** is found where Str1 = QuList[ite] and Str2 = Inpat[ite], ite = 1, 2, 3, 4 and the resultant LCS is matched with ***'Str2'***. If the resultant LCS gets

a match with ***'Str2'*** then this module returns 1, otherwise 0. The return value of this module is stored in Param2. If Param2 = 1 then, Param4 is incremented by 1. This process is done for each item present in the list ***'MatchFirst'***. If Param4 = 4 for an item in ***'MatchFirst'*** then that file name is copied to the list ***'MatchSecond'***. According to the research, the ***'MatchCommonFeature'*** selects the correct match from the multiple matches in the list ***'MatchFirst'***. But if for some images, both the modules of ***'Recognition'*** result in multiple matches or no matches, then ***'TraceAnotherDirection'*** module is called.

c) TraceAnotherDirection Module

This module consists of two parts '***DictionaryAnotherWay'*** and ***'FindingMatchAnother'***. For extraction of features from the dictionary of images stored in the directory ***'Dictionary'***, the ***'DictionaryAnotherWay'*** undergoes same modules as in ***'DictionaryBuilding'***, the only difference being the direction of tracing of the continuous black pixels in the '***FindPath'*** module in ***'DictionaryBuilding'*** for the four quadrants. The modified direction of the tracing of the continuous black pixels is shown in a module named ***'FindPathAnother'***.

row = Number of rows of quadrant

col = Number of columns of quadrant

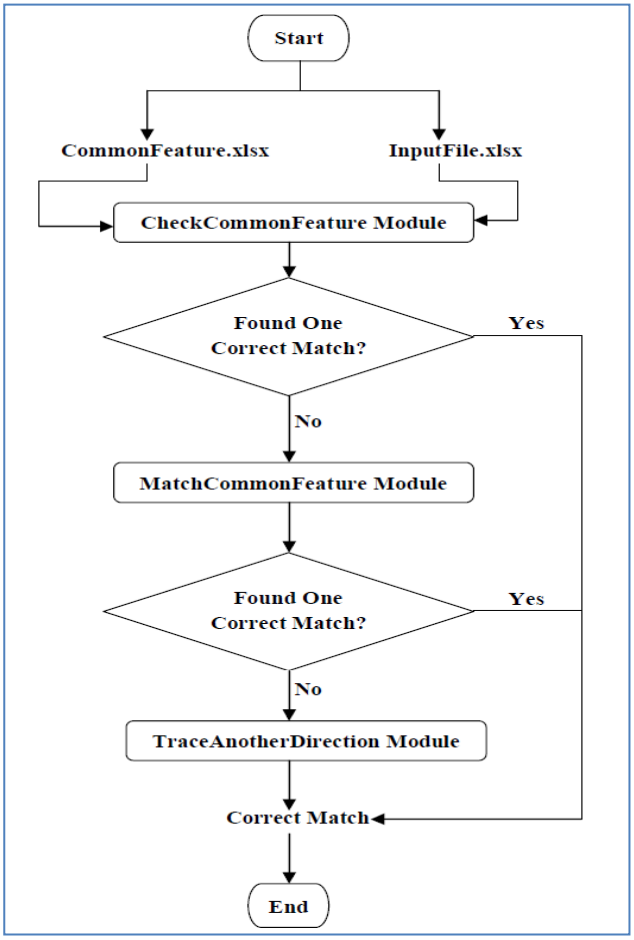


Fig. 6. Recognition of Input Image.

Algorithm:

FindPathAnother(quadNo, quadrant, DicItem, DicInnerItem, DataPath, shortName)

1. SET I = 0
2. SET J = col – 1
3. IF quadNo = 1 THEN GO TO STEP 4
4. REPEAT STEP 5 WHILE J > 0
5. REPEAT STEP 6 WHILE I < row
6. IF quadrant[I][J] = 0 THEN GO TO STEP 7
7. CALL GettingFeaturesLeft(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
8. SET I = row – 1
9. SET J = col – 1
10. IF quadNo = 2 THEN GO TO STEP 11
11. REPEAT STEP 12 WHILE I > 0
12. REPEAT STEP 13 WHILE J > 0
13. IF quadrant[I][J] = 0 THEN GO TO STEP 14
14. CALL GettingFeaturesLeft(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
15. SET I = 0
16. SET J = 0
17. IF quadNo = 3 THEN GO TO STEP 18
18. REPEAT STEP 19 WHILE I < row
19. REPEAT STEP 20 WHILE J < col
20. IF quadrant[I][J] = 0 THEN GO TO STEP 21
21. CALL GettingFeaturesRight(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
22. SET I = 0
23. SET J = col – 1
24. IF quadNo = 4 THEN GO TO STEP 25
25. REPEAT STEP 26 WHILE I < row
26. REPEAT STEP 27 WHILE J > 0
27. IF quadrant[I][J] = 0 THEN GO TO STEP 28
28. CALL GettingFeaturesLeft(I, J, quadrant, quadNo, DicItem, DicInnerItem, DataPath, shortName)
29. EXIT

'FindingMatchAnother' finds a correct match from the '***CommonFeature2.xlsx***'(output of ***'DictionaryAnotherWay').*** For Feature extraction, the input image present in the directory ***'Input'*** undergoes same modules as in ***'FindingMatch'***, except the direction of tracing of the continuous black pixels in the '***FindPath'*** module in ***'FindingMatch'*** for the four quadrants. The modified direction of the tracing of the continuous black pixels is shown in a module named ***'FindPathAnother'***. The same modules of ***'Recognition'*** are used for finding a correct match. As per the research, the ***'TraceAnotherDirection'*** gives a correct match for the input image.

IV. Results

This paper deals with recognising a printed Odia alphabet in an image which is created by scanning a document or document converted to image by using a software, both

written in a font family ***'AkritiOriAshok-99'*** in a particular font size. The font sizes that have been considered are 18, 20, 22, 24, 26, 28, 32, 48, and 72.

To achieve recognition of an Odia alphabet, the system explained in this paper is divided into two parts; one is ***'DictionaryBuilding'*** and other is ***'FindingMatch'.*** The ***'DictionaryBuilding'*** takes a directory ***'Dictionary'*** (consisting of images of Odia alphabet), undergoes through several modules to extract features from images present in ***'Dictionary'*** and the extracted features are written in an excel file, ***'DictionaryFeatures.xlsx'***. The common feature found out from the extracted features in all font sizes for each Odia alphabet (***'DictionaryFeatures.xlsx'***) is written in an excel file, ***'CommonFeature.xlsx'***. The ***'FindingMatch'*** takes an image of Odia alphabet as input and the alphabet can be in any font size of font family ***'AkrutiOriAshok-99'***; features are extracted from the input image and the extracted feature is given as input to ***'Recognition'***. The Recognition finds a correct match for the input image.

The Lenovo ideapad 310 Laptop with 64-bit Windows 10 Operating system, 4GB RAM and Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.70 GHz have been used for the system. The JetBrains PyCharm Community Edition 2019.1 as Integrated Development Environment (IDE) and opencv-python 4.1.1.26 libraries has been used to implement the system.

For testing, an image of Odia alphabet is given as input to the ***'FindingMatch'*** to find a correct match. Nine font sizes have been considered for this research and 200 images of Odia alphabet of each font size making a total of 1800 images are provided as input to ***'FindingMatch'*** one at a time. The percentage of correctness has been shown in Fig. 7.

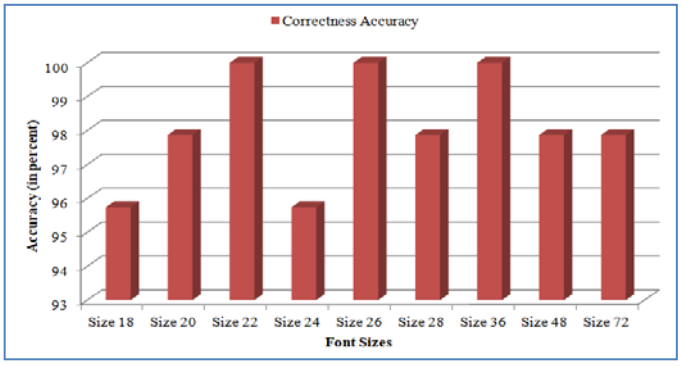


Fig. 7. Correctness Accuracy of Odia Alphabets in Different Font Sizes.

For feature extraction, [22] and [23] had used Water-Reservoir Principle to get the shapes of the characters and numerals respectively; [24] had divided the characters into nine zones and traced the shapes in each zone; [25] had found out the centroid of the character and then the angle between the centroid and the pixel to trace the shapes of the characters; and [26] had first found out some low-level strokes to detect the high-level strokes and using these strokes, the shapes of the character had been traced. The proposed approach has also traced the shapes of the characters by first dividing the character into four quadrants and then scanning each quadrant in different directions to get the features in string format. The proposed system has also been compared with the systems in [22], [23], [24], [25] and [26], and the results have been tabulated in the Table I.

TABLE I. ACCURACY COMPARISON OF PROPOSED APPROACH WITH OTHER APPROACHES

Accuracy Achieved by the Approaches in Related Work				**Accuracy Achieved by the Approach in this Paper**
Approaches	**Language**		**Accuracy**	
[22]	Handwritten Odia	Isolated Characters	98.6%	98.1%
		Two-Character Touching Components	96.7%	
		Three-Character Touching Components	95.1%	
[23]	Handwritten English Numerals		94.8%	
[24]	Printed Odia		92%	
[25]	Printed Odia		91.3%	
[26]	Printed Gujarati		96.87%	

It has also been found that alphabet Chota U (ଉ) is recognised as Bada U (ଊ) in some font sizes because they have very little difference in their structure. The system faces the same challenge for the alphabets Ra (ର) and Ru (ଋ).

V. CONCLUSION

The approach described in this paper goes through two parts. First part deals with building a dictionary and the second part deals with finding a match for the image given as input. In the first part (***'DictionaryBuilding'***), a dictionary of images consisting of alphabets in the font family ***'AkrutiOriAshok-99'*** and in different font sizes are prepared. Then features are extracted from the images and written in an excel file, ***'DictionaryFeatures.xlsx'***. LCS has been used to find the common feature from the extracted features and the common feature has been written in an excel file, ***'CommonFeature.xlsx'***. The second part deals with finding a match for the image that is given as input. In second part, features are extracted from the input image and matched with the feature present in ***'CommonFeature.xlsx'***. In some cases, if more than one match or no match is found then the four quadrants of the input image have been scanned in another direction. The overall correctness accuracy of the system has been achieved as 98.1%.

As the proposed approach recognises Chota U (ଉ) as Bada U (ଊ) and Ra (ର) as Ru (ଋ) in some font sizes, hence, further research can be done in future to eliminate this disadvantage. Elimination of this problem may increase the accuracy percentage. Moreover, research can be done to reduce the number of phases of the proposed system which may increase the efficiency of the system.

REFERENCES

[1] Devabrata Kar, Chabila Madhu Barnabodha, Published by Odisha Book Emporium.

[2] Pandit Narayan Mohapatra, Sridhar Das, Sarbasara Byakarana, ISBN: 8186085009, Published by New Students' Store.

[3] Aradhana Kar, Sateesh Kumar Pradhan, "A Three-Phase Noise Removal Approach to Achieve Accuracy in Line Segmentation of Odia text", 19th OITS International Conference on Information Technology (OCIT), pp. 54-59, 2021.

[4] "Image Thresholding, Image Processing in OpenCV" (Web Search)

[5] Ravishankar Chityala, Sridevi PudiPeddi, Image processing and Aquisition using Python", CRC Press Taylor & Francis Group, Chapman & Hall, pp 141 – 143.

[6] Nobuyuki Ostu, "A Threshold Selection Method from Gray-Level Histograms", IEEE Transactions on Systems, Man, and Cybernetics, Volume 9, Issue 1, pp 62 – 66, January 1979.

[7] S. Sridhar, "Digital Image Processing", Oxford University Press, pp 10 – 11, 2013.

[8] Rafael C. Gonzalez, Richard E. Woods, Digital Image Processing, Pearson, pp 55 – 65, Third Edition.

[9] Mark S. Nixon, Alberto S. Aguado, Feature Extraction & Image Processing for Computer Vision, Elsevier, Academic Press, pp 37 – 41, Third Edition.

[10] "os.path", https://docs.python.org/3/library/os.path.html

[11] Eric Gazoni, Charlie Clark, "openpyxl - A Python library to read/write Excel 2010 xlsx/xlsm files", Version 3.0.10.

[12] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein, "Introduction to Algorithms", The MIT Press, Tata McGraw Hill Book Company, pp 350 – 355, Second Edition.

[13] Udit Agarwal, "Algorithms Design and Analysis", Dhanpat Rai & Co (P) Ltd, pp 262 – 270, Second Edition.

[14] Narasimha Karumanchi, "Data Structures and Algorithms Made Easy", CareerMonk Publications, pp 373 – 375.

[15] Steven S. Skiena, "The Algorithm Design Manual", Springer, Second Edition, pp 650 – 653.

[16] Lekh Raj Vermani, Shalini Vermani, "An Elementary Approach to Design and Analysis of Algorithms", Primers in Electronics and Computer Science, Vol. 4, World Scientific, pp 174 – 185, 2019.

[17] Jan Erik Solem, "Programming Computer Vision with Python: Tools and Algorithms for Analyzing Images", O'Reilly, pp 7 – 8.

[18] "NumPy" (Web Search)

[19] "Matplotlib.pyplot" (Web Search)

[20] Zed A. Shaw, "Learn Python 3 the Hard Way A Very Simple Introduction to the Terrifying Beautiful World of Computers and Code", Addison-Wesley, Exercise 34 (Lists), pp 120 – 121 and Exercise 39 (Dictionaries), pp 140 – 144.

[21] Kent D. Lee, "Python Programming Fundamentals", Springer, 2014.

[22] N. Tripathy, U. Pal, "HandWriting Segmentation of Unconstrained Oriya Text", Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition, IEEE Computer Society, 2004.

[23] U. Pal, A. Belaid, Ch. Choisy, "Touching numeral segmentation using water reservoir concept", Pattern Recognition Letters 24, pp 261-272, 2003.

[24] Dibyasundar Das, Ratnakar Dash and Banshidhar Majhi, "Odia Compound Character Recognition Using Stroke Analysis", Computational Intelligence in Data Mining, Advances in Intelligent Systems and Computing, volume 556, pp 325 – 332, 2017.

[25] Debananda Padhi, Debabrata Senapati, "Zone Centroid Distance and Standard Deviation Based Feature Matrix for Odia Handwritten Character Recognition", Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), Springer-Verlag Berlin Heidelberg, pp 649 – 658, 2013.

[26] Mukesh M. Goswami, Suman K. Mitra, "Printed Gujarati Character Classification Using High-Level Strokes", Proceedings of 2nd International Conference on Computer Vision & Image Processing (Volume 2), Advances in Intelligent Systems and Computing (Volume 704), Springer, pp 197 – 209, 2017.

[27] Mukesh M.Goswami, Suman K. Mitra, "Classification of Printed Gujarati Characters using Low-Level Stroke Features", ACM Transactions on Asian and Low-Resource Language Information Processing, Volume 15, Issue 4, Article 25, pp 1 – 26, June 2016.

Enhancement of Design Level Class Decomposition using Evaluation Process

Bayu Priyambadha[1], Tetsuro Katayama[2]
Interdisciplinary Graduate School of Agriculture and Engineering, University of Miyazaki, Miyazaki, Japan[1, 2]
Faculty of Computer Science, Universitas Brawijaya, Malang, Jawa Timur, Indonesia[1]

***Abstract*—Refactoring on the design level artifact such as the class diagram was already done using the threshold-based agglomerative hierarchical clustering method, specifically class decomposition. The approach produced a better cluster based on the label name similarity of attribute and method. But, some problems emerge from the experiment result. The negative Silhouettes element still exist in the cluster. And, there is an unusable cluster that only consists of one attribute element. This paper has proposed the evaluation process to optimize the result of clustering. This evaluation process is an additional process that aims to move the negative Silhouettes element to the other cluster. The movement is also to get the better value of element Silhouettes value. The evaluation process can produce a better result for clusters. The clusters produced from the evaluation process have higher Silhouettes values. The average Silhouettes value is increased by about 40%. Ultimately, the result shows no unusable cluster as mentioned in the previous research.**

Keywords*—Refactoring; design level refactoring; software refactoring; class decomposition; software quality

I. Introduction

Refactoring the software design artifact is essential to maintain the design's internal structure [1]. Changes or alteration to the design artifact is easier than source code. The easiness is because of the original character of the model. The model is easy to change because the model is an abstract description of something more detailed. Refactoring at the design level means the refactoring activity is using the software design artifact as an object. The easiness of alteration and simulation of quality measurement using the specific metric is one of the benefits of refactoring software design artifacts. The software design artifact is a model bridging the requirement and implementation artifact and is the center of the software development process [2]. The changes will impact both sides, requirement and implementation artifact. In the case of software maintenance, refactoring activity is one way to decrease the maintenance cost [3], [4]. In the case of software development, the refactoring activity can be used as the evaluation process to maintain the internal quality of the design artifact before it is implemented into the source code. The design level refactoring increases the quality awareness of the design artifact as early as possible.

Shifting the object of refactoring activity to a higher level of abstraction has a specific problem. There is a limitation of information in the design artifact rather than the implementation artifact [5]. Therefore, excavation or mining and in-depth information analysis of the design artifacts are necessary [6]–[8]. Generally, the information on the design artifact is only written on the artifact. Sometimes, the information contains a hidden meaning that needs extra analysis to mean it. Natural language processing or semantic analysis is one approach that provides the functionality to gain the meaning of information [7]. Different from that, the source code level information clearly defines a specific element. The software developer can easily use it as data to analyze and measure quality, for example, the number of operand or operators in the source code to measure the complexity of the source code. The developer also can easily know the relation between attribute and method by reading the internal source code. They can figure it out by looking at the assigning value statement to the specific attribute.

Refactoring activity begins from the existence of the smell in the software artifacts. The smell is the indicator that there is something wrong in it. The quality of the artifacts decayed because of the smell. Finding the smell in the artifact is the first activity before refactoring itself. Researchers have already researched the smell detection process in the software artifacts. Smell detection mostly uses the source code as an object, known as code smell detection. Nowadays, the design of smell detection has started to emerge [9]–[11]. The terms and characteristics between code and design smells are different. The differentiation is based on the object, and the information lies in it. But, the previous research tried to use the code smell term and characteristics to find the smell in the design artifacts [8]. As a result, the Blob smell is detected using the class diagram information. Semantic analysis was used to determine the relation between class elements based on the name labels to enrich the class diagram information.

Blob smell is one of the lacks of internal structure quality indicators. It indicates the greedy process of the class. One class has a lot of process in it, whereas the other class nearby is only the data provider. The blob class monopolizes data processing provided by the nearby classes [8]. This phenomenon can happen due to software changes or the developer's lack of clean architecture theory. The clean architecture theory explains that the class must comply with the Single Responsibility Principle [12]. During the development or evolution cycle, the class has to have only one reason to change. The reason to change is related to the process or functionality of the specific class. If the class has more than one function or manipulates many operations, it will be the candidate that there is much reason to change it during the software cycle.

Furthermore, the blob class in the software system will increase the maintenance cost because it affects the class's

understandability [13]. The refactoring activity can resolve the problems of the Blob class. One of the refactoring processes to solve the blob class problem is class decomposition. Mostly, a source code became the field of the class decomposition process. Much research has been conducted using source code information on the class decomposition process. Shifting the decomposition process to the design level is interesting due to the possibility of decreasing the cost of the change, easiness of change simulation, and early quality awareness.

Class decomposition is the process of separating one class into many classes. The decomposition is based on the specific characteristics defined before the separation process. Many researchers proposed the class decomposition mechanism at the source code or design level. Most of the class decomposition mechanism approach uses the clustering process [13]–[19], in which the elements in the class are separate based on each element's closeness characteristic. This method aims to make the separation process result following the Single Responsibility Principle concept. The Threshold-based Agglomerative Hierarchical Clustering was tried to implement on both source code [19] and design level artifacts [5]. Each clustering process is based on the metrics generated from each source code and design artifacts. The design level class decomposition on the class diagram uses two metrics, syntactic (syn) and semantic (sem) [5]. Both metrics are calculated by considering syntactic and semantic closeness from the element's label name. Using the closeness of syntactic and semantic of the label name, the Threshold-based Agglomerative Hierarchical Clustering created a more compact cluster compared to the result of clustering on the source code level. The compactness of clusters was observed from the value of the Silhouettes value of every cluster. However, the decomposition results still show the elements with a negative Silhouette value. A negative Silhouette value indicates that an element's distance from the others in its cluster is large. The negative Silhouette elements are considered the worst in the relation with the concept of single responsibility principle. It is important to enhance the element placement mechanism of the negative element. Additionally, some clusters are considered unable to implement because, in case implemented as a class, it will instantiate objects that cannot interact with each other. A cluster with only one element, especially if the element includes a private modifier, is deemed worthless or useless. As a result, it is seen to be critical to incorporate the modifier aspect in the decomposition process.

The validity of the class decomposition's result is important. It is related to the class's applicability when implemented in the real case or source code. The existence of negative elements in the resulting cluster and a single private element in one cluster is a big problem for the applicability of the class. This condition requires in-depth attention, especially to validate the result of the decomposition process. The basic validation mechanism is to move the specific element from the origin cluster to the other cluster. The moving mechanism aims to put the specific element (negative element) to the other cluster to get a better Silhouette value. The other problem is the existence of the private single-member cluster. It also decreases the applicability of the class when it is implemented into the software's source code. In the previous approach, the clustering process is based on the two metrics syn and sem. The addition of other metrics is important to solve the unusable class.

This research is conducted to propose the validation mechanism to solve previous research's problems. The basic idea is to move the elements in the cluster that are not well-positioned. The new metric is proposed to increase every cluster's placement accuracy and compactness. All descriptions of the proposed algorithm of the validation mechanism and the experiment are organized as follows. Section II summarizes the state of the arts of the class decomposition approach. Then continue the description of the class usability and compactness of the class in the decomposition process in Section III. Section IV and V explain the proposed algorithm and the research scenarios. Section VI describes the result and discussion. Then the last is the conclusion and future work in Section VII.

II. Related Work

Many researchers published methods for class decomposition based on a specific type of smell. The research has two object studies, source code, and class diagram. The following section summarizes the history of research in the area of class decomposition.

A. Class Decomposition on Source Code

Bavota et al. presented a number of methods that could be used to decompose classes at the level of source code. Bavota's research history used the two-step decomposition techniques and MaxFlow-MinCut algorithms to extract classes [14]–[16]. The research involved considering both structural and semantic characteristics of the class. There are three metrics used: Structural Similarity between Methods (SSM), Call-based Interaction between Methods (CIM), and Conceptual Similarity between Methods (CSM). According to a study, transitive closure was calculated using metrics based on the values of distance between class elements. The other method uses the graph to represent the relatedness between elements and the weight to represent the closeness between elements. The transitive closure is able to split a Blob class into more than two classes, which is a significant improvement over the MaxFlow-MinCut approach. Furthermore, it can automatically determine how many classes should be extracted from a Blob.

A discussion of metric-based refactoring opportunities identification for object-oriented software systems is presented in an article by Isong Bassey et al. [20]. They conducted a thorough analysis of sixteen (16) primary studies in order to identify the state of the practice in ROI. The purpose of this article is to summarize all existing refactoring opportunities. The analysis was divided into three categories: structural, semantic, and structural and semantic. Using metrics to identify refactoring opportunities is the focus of this paper. Al Dallal's structural approach and Bavota's structural and semantic approaches previously published elsewhere are summarized in this paper.

According to Wang Ying et al., weighted clustering is automatically used to refactor software [13]. This article focuses on class-level refactoring. A network is considered to be a representation of the relationship of dependencies between

methods (as nodes). There are three matrices that illustrate the relationships between methods, (i) attribute sharing (Sharing Attribute Weight/ SAW), (ii) method invocation (Method Invocation Weight/ MIW), and (iii) functional coupling (Functional Coupling Weight/ FCW). A combination of three matrices as well as semantic similarity weights (SSW) is used to compute edge weights. Thus, the most advantageous cluster with the appropriate weight is selected. Wang compares his method with Bavota and Fokaefs. Wang's approach improves cohesion and coupling without affecting the code's behavior. Furthermore, it improves code's understandability, flexibility, reusability, and maintainability.

Mohamed Hamdi discusses the Agglomerative Hierarchical Clustering (AHC) method for class decomposition [19]. The decomposition occurs until classes have a single responsibility iteratively. One of the main challenges is terminating the decomposition process. They define the threshold concept for determining the endpoint during the decomposition process. There are six matrices: Internal Attribute Sharing (IAS), Internal Direct Call Dependency (IDC), Internal Indirect Call Dependency (IIC), Internal Method Sharing (IMS), and External Indirect Call Dependency (EIC). In this case, the weighted AHC results are more beneficial. This approach appeared to be a solution to the problem of the limited number of classes resulting from the decomposition process and the termination state.

B. Model-Driven Software Engineering

Model-Driven Software Engineering (MDSE) uses a software model as the primary artifact of software development [2]. Compared to the implementation artifact (source code), the software model is closer to the problem domain. The model transformation is the heart of the MDSE since the MDSE aims to generate the source code from the models. On the other hand, there is another approach to the development of software called Code-centric Development (CcD). A comparison study between MDSE and CcD has already been done for over a decade. From the review paper by Domingo et al., many researchers have been evaluating the benefit of the MDSE [21]. Some works said that MDSE decreases development time (up to 89%) relative to Code-centric Development (CcD). The other works suggest that the MDSE is suitable for academic exercise. Furthermore, the other works assert that MDSE is also suitable for inexperienced developers. Finally, Domingo et al., based on their review of MDSE, conclude that the MDSE is suitable for academic exercise and inexperienced developers [21].

C. Class Decomposition on Class Diagram

The class decomposition process is shifted to the design level artifact taking into account the ease of change and quality measurement. A similarity score is calculated between the class's elements (attributes and methods) used in the decomposition process based on the metrics that are derived from the information found in the class diagram. There are two approaches to determining the similarity rates between elements of a class: syntactic and semantic analysis. Thus, the two approaches evaluate the similarity of sentences based on their similarity in terms of syntax and meaning. Those metrics are syn and sem. The following formulas are defined for the metrics [5].

$$syn = \begin{cases} 1, & similar\ type > 0 \\ 0, & similar\ type \leq 0 \end{cases} \quad (1)$$

and,

$$sem = \frac{2.wi.|w_1 \cap w_2| + ws.(|s(w_1,w_2)| + |s(w_2,w_1)|)}{|w_1| + |w_2|} \quad (2)$$

where $s(w_1, w_2)$ or $s(w_2, w_1)$ is the number of words that have a synonym relationship between two labels, and $wi = 1$ and $ws = 0.75$ [22]. The closeness or similarity between class elements is calculated using the following formula.

$$Sim(e1, e2) = \frac{syn + sem}{2} \quad (3)$$

The class decomposition process uses the Threshold-based AHC that is used the similarity formula to calculate the closeness between class elements. Based on the previous decomposition result, the static and dynamic threshold AHC clusters are more compact than Hamdi's approach. The compactness of the clusters is measured using the Silhouette value. Based on the results, it is evident that there are certain advantages to be gained, but there are also some shortcomings as well. Decomposition results still show elements with negative Silhouette values. When the Silhouette value is negative, it indicates that the current element is far from the other elements in the cluster. In other words, negative Silhouette elements are considered to be the worst. Negative elements need to be improved in their movement mechanism. Moreover, some clusters are considered unable to implement because their objects may not be able to collaborate. It is considered useless to have a cluster with only one element, especially if the element has a private modifier. This is why the modifier aspect must be included in the decomposition process. Avoiding useless clusters is essential.

III. Proposed Approach

A. Scope of Study

The research focuses on our previous research results using the same dataset as the previous experiment. Two formulas will be proposed to solve the previous problems. Those formulas will be focused on overcoming the negative Silhouette and useless cluster [5] by combining Class Usability ($CUsability$) and Silhouettes value ($s(i)$). The combination of two metrics are used to evaluate the cluster after the clustering process. The whole evaluation process will be proposed as an evaluation algorithm. This study also uses classes that are not problematic to gain other insights in this study.

B. Problem Accomplishment

The previous research's result mentioned that there were two problems found. The first is the negative Silhouette element, and the second is the cluster that is predicted to be unusable. That is why the evaluation process must consider two aspects: the Silhouette value and the usability of the class

(that will be quantified in the form of a metric). The following formula calculates the Eval value ($Eval$).

$$Eval = a.s(i) + b.CUsability \quad (4)$$

$s(i)$ shows the Silhouette value, and $CUsability$shows the class usability value. a and b is the weight that describes the proportion of each value to the evaluation value. $CUsability$ has measured the usability of the class by considering the number of public methods that existed in the cluster. The class is usable if it at least consists of one public method. In other words, if the class has a public method, the class will be able to collaborate with the others (useful). The $CUsability$ is calculated using the following formula.

$$CUsability = \begin{cases} 0 & ,mpub = 0 \\ 1 & ,mpub \geq 1 \end{cases} \quad (5)$$

The $mpub$ is the number of public methods in the class candidate (in the cluster). The silhouette value is calculated using the following formula [23].

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (6)$$

Where,

- $a(i)$ = the average dissimilarity of i to all other objects of A, then,
- $d(i, K)$ = the average dissimilarity of i to all object cluster K, when $K \in Cluster\ and\ K \neq A$,
- $b(i) = MIN\big(d(i, K)\big), K \neq A$.

The proposed evaluation algorithm has the main process of selecting the negative element and then moving it to the other cluster by comparing the Eval value before and after moving. The algorithm will be appended to the previous algorithm as the evaluation process.

C. Preliminary Experiment

Before the algorithm is confirmed, a preliminary experiment is conducted to ensure the performance of the evaluation process. The preliminary experiment uses one study case from the Landfill dataset, Class Transfer (Blob class from HSQLDB). The preliminary experiment is an early evaluation of the proposed algorithm (implementation of the $Eval$ formula) that uses a combination of weights a and b. In the case of Class Transfer, using a combination of weights with the value a bigger than b in the $Eval$ formula, the process was always run because the negative element always existed. As a result, the process of evaluation is never stopped. Based on this result, it was tried to print out the difference of $Eval$ value (before and after moving) every iteration and draw it into the line graph to show the trend. Fig. 1 shows the line graph of $Eval$ value differences in the case Class Transfer using weights $a = 0.9$ and $b = 0.1$. The trend shows that the values are shifted alternately in the mid to late iterations. It seems that the specific element was moved and moved back to the same cluster because the value of Silhouettes of that particular element is always negative in both clusters (origin and destination cluster).

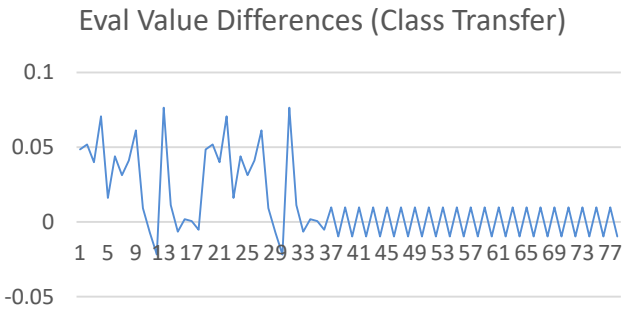


Fig. 1. Eval Value Differences of Class Transfer.

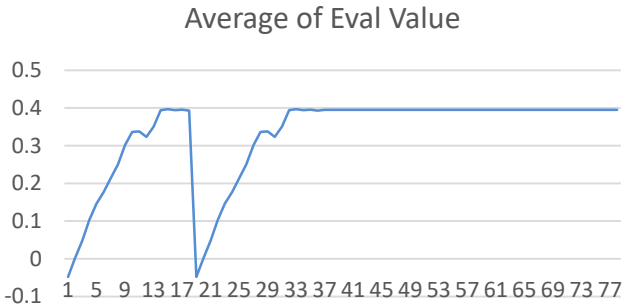


Fig. 2. The Average of Eval Value.

Fig. 1 shows the graph of the Eval value in every iteration. In the middle of the graph, Fig. 1, the data show a pattern that causes the unstoppable process. Even though it shows the pattern, the data seems unstable (continually moving from positive to negative). So, it needs to calculate to get a more stable value. Starting from iteration number 37, the Eval value between before and after moving is 0.00977 and -0.00977. Then, it tried to use the average formula to get a more stable value.

Fig. 2 shows the result after the values are averaged. The graph shows the flat value starting from iteration number 37, and it is easier to use as a termination condition for the algorithm for Class Transfer. The flat value of Class Transfer is about 0.4. The value of 0.4 cannot be used in the other study cases. It is only suitable for Class Transfer. Therefore, it is possible to be different from the other study cases. The other calculation is necessary to find a universal value to get the stopping condition.

D. Stopping Condition of Algorithm

The stopping condition in the decomposition process was the new problem that emerged in the preliminary experiment. One formula that can be used to find the pattern is by calculating the average Eval value between pre and post-movement to the other cluster. The following formula represents how the average of $Eval$ can be calculated.

$$AvgEval_n = \frac{Eval_n + Eval_{n-1}}{2} \quad (7)$$

Where,

- n is the number of decomposition iterations,
- $Eval_{n-1}$ is Eval value before moving to the other cluster,
- $Eval_n$ is Eval value after moving to the other cluster.

The $AvgEval$ for every iteration is represented in a line graph in Fig. 2. The easiest way to find the stopping condition is to calculate the differences of $AvgEval$ between two iterations using the following formula.

$$AvgDiff = AvgEval_n - AvgEval_{n-1} \quad (8)$$

Then the stop condition is represented as follows.

$$StopCondition = \begin{cases} AvgDiff = 0 & ,true \\ AvgDiff \neq 0 & ,false \end{cases} \quad (9)$$

Where,

- $AvgEval_n$ is Average Eval value from iteration number n,
- $AvgEval_{n-1}$ is Average Eval value from iteration number $n-1$.

The main idea of the stopping condition is to find zero (0) differences of $AvgEval$ between iterations. If the differences of $AvgEval$ are zero (0), then it means that there is no increment of Eval value even if the specific element is moved to the other cluster. Then the last position of the cluster will be chosen as the best solution. Fig. 3 shows the line graph of $AvgDiff$ as the representation of the differences of $AvgEval$ before and after movement.

E. Proposed Algorithm

The algorithm is proposed to answer the problems that emphasize the previous class decomposition approach. The new algorithm is the representation of an additional process on the class decomposition. The evaluation algorithm is described in Fig. 4. In the design level class decomposition research, the decomposition process is done by the Threshold-based Hierarchical Agglomerative Clustering. First, two metrics are used to calculate the closeness between elements in the class decomposition process, syn, and sem. Then, the process continues to the evaluation process. That is aim is to evaluate the placement of every element. This process is focused on the element that has the negative silhouette value. The evaluation process's main idea is to move the negative element to the other cluster to get a better silhouette value. The evaluation process is an iteration process that considers the increment of silhouette value and stopping condition that is defined in the previous section.

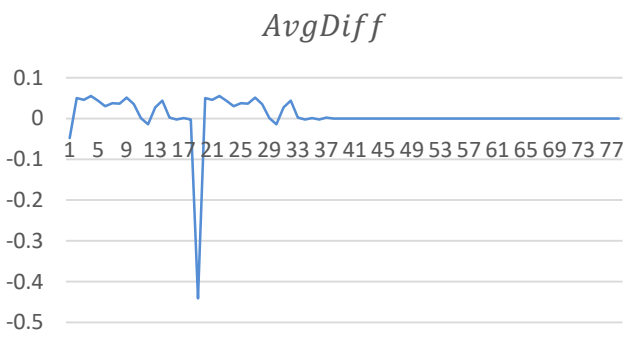


Fig. 3. The Line Graph $AvgDiff$

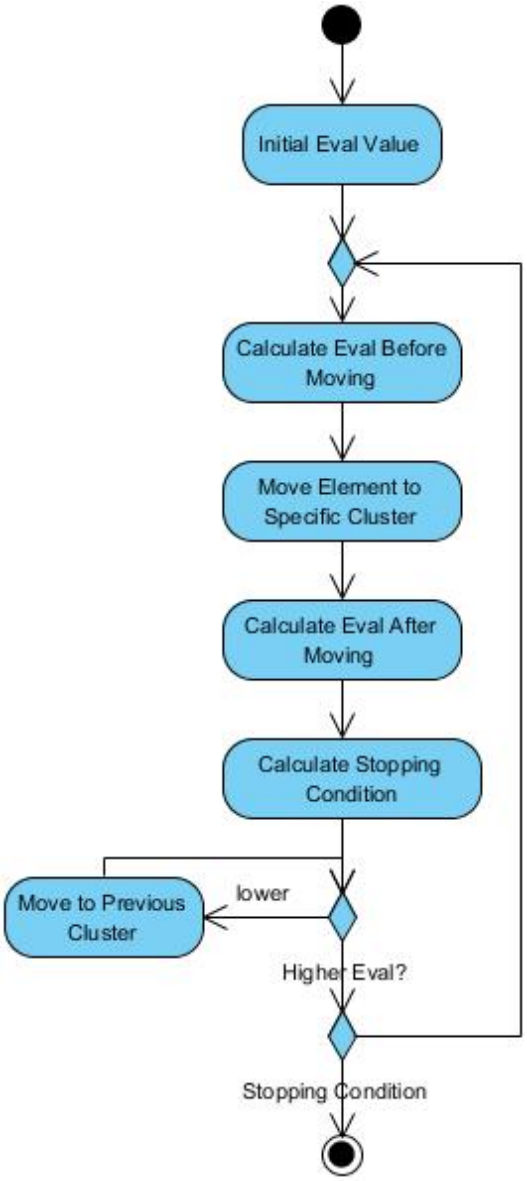


Fig. 4. The Evaluation Algorithm

IV. The Experiment Scenario

The class decomposition experiment uses ten study cases taken from the open-source Java application. The test cases are classes in several open-source Java application that is indicated as Blob smell according to the Landfill smell dataset [24]. Table I shows the list of test cases. All test cases will be decomposed using the threshold-based hierarchical agglomerative clustering and static and dynamic thresholds. Then evaluate using the proposed approach in various combinations of weights. The weights are set to start from a=0.1 and b=0.9 until a=0.9 and b=0.1, with the increment and decrement of 0.1.

The combination of weights has aimed to find the best composition of weights in the class decomposition process based on the compactness and usability of clusters.

TABLE I. List of Test Cases

No.	Class Name	Application
1.	AudioFile	aTunes
2.	JDBCBench	HSQLDB
3.	Interpreter	jEdit
4.	SVGOutputFormat	jHotDraw
5.	Transfer	HSQLDB
6.	Import	agroUML
7.	StringConverter	HSQLDB
8.	RipCdDialog	aTunes
9.	DefaultDrawingViewTransferHandle	jHotDraw
10.	MDIApplication	jHotDraw

TABLE II. The Result of Static Threshold Decomposition

Class	Threshold	a=0.1;b=0.9	a=0.2;b=0.8	a=0.3;b=0.7	a=0.4;b=0.6	a=0.5;b=0.5	a=0.6;b=0.4	a=0.7;b=0.3	a=0.8;b=0.2	a=0.9;b=0.1
AudioFile	**Clusters**	2	2	2	2	2	2	2	2	2
	Eval	0.932	0.864	0.796	0.728	0.661	0.593	0.525	0.457	0.389
	Silhouettes	0.322	0.322	0.322	0.322	0.322	0.322	0.322	0.322	0.322
JDBCBench	**Clusters**	2	2	2	2	2	2	2	2	2
	Eval	0.935	0.871	0.807	0.743	0.679	0.615	0.551	0.487	0.423
	Silhouettes	0.359	0.359	0.359	0.359	0.359	0.359	0.359	0.359	0.359
Interpreter	**Clusters**	1	1	1	1	1	1	2	2	2
	Eval	0.928	0.856	0.784	0.712	0.64	0.569	0.436	0.372	0.309
	Silhouettes	0.281	0.281	0.281	0.281	0.281	0.281	0.246	0.246	0.246
SVGOutputFormat	**Clusters**	2	2	2	2	2	2	2	2	2
	Eval	0.919	0.839	0.759	0.679	0.599	0.519	0.439	0.359	0.279
	Silhouettes	0.199	0.199	0.199	0.199	0.199	0.199	0.199	0.199	0.199
Transfer	**Clusters**	1	1	1	1	1	2	2	2	2
	Eval	0.929	0.859	0.788	0.718	0.648	0.418	0.266	0.225	0.258
	Silhouettes	0.296	0.296	0.296	0.296	0.296	0.216	0.179	0.228	0.291
Import	**Clusters**	2	2	2	2	2	2	2	2	2
	Eval	0.917	0.835	0.753	0.671	0.588	0.506	0.424	0.342	0.26
	Silhouettes	0.177	0.177	0.177	0.177	0.177	0.177	0.177	0.177	0.177
StringConverter	**Clusters**	2	2	2	2	2	2	2	2	2
	Eval	0.952	0.904	0.856	0.808	0.76	0.712	0.664	0.616	0.569
	Silhouettes	0.521	0.521	0.521	0.521	0.521	0.521	0.521	0.521	0.521
RipCdDialog	**Clusters**	2	2	2	2	2	2	2	2	2
	Eval	0.876	0.807	0.739	0.67	0.602	0.534	0.443	0.388	0.292
	Silhouettes	0.26	0.26	0.26	0.26	0.26	0.26	0.276	0.276	0.238
DefaultDrawingView TransferHandle	**Clusters**	2	2	2	2	2	2	2	2	2
	Eval	0.93	0.86	0.791	0.721	0.652	0.582	0.513	0.443	0.374
	Silhouettes	0.304	0.304	0.304	0.304	0.304	0.304	0.304	0.304	0.304
MDIApplication	**Clusters**	2	2	2	2	2	2	2	2	2
	Eval	0.927	0.854	0.782	0.709	0.637	0.564	0.492	0.419	0.347
	Silhouettes	0.274	0.274	0.274	0.274	0.274	0.274	0.274	0.274	0.274

TABLE III. The Result of Dynamic Threshold Decomposition

Class	Threshold	a=0.1;b=0.9	a=0.2;b=0.8	a=0.3;b=0.7	a=0.4;b=0.6	a=0.5;b=0.5	a=0.6;b=0.4	a=0.7;b=0.3	a=0.8;b=0.2	a=0.9;b=0.1
AudioFile	**Clusters**	7	7	7	7	7	7	7	9	7
	Eval	0.951	0.903	0.855	0.807	0.759	0.711	0.663	0.404	0.567
	Silhouettes	0.519	0.519	0.519	0.519	0.519	0.519	0.519	0.28	0.519
JDBCBench	**Clusters**	2	2	2	2	2	6	6	6	7
	Eval	0.935	0.871	0.807	0.742	0.678	0.361	0.266	0.239	0.266
	Silhouettes	0.357	0.357	0.357	0.357	0.357	0.016	0.185	0.23	0.292
Interpreter	**Clusters**	1	1	1	1	6	5	11	9	6
	Eval	0.928	0.856	0.758	0.712	0.413	0.402	0.372	0.352	0.293
	Silhouettes	0.281	0.281	0.281	0.281	-0.013	0.138	0.154	0.31	0.268
SVGOutputFormat	**Clusters**	5	5	5	5	10	9	9	8	6
	Eval	0.936	0.873	0.81	0.746	0.46	0.418	0.345	0.4	0.44
	Silhouettes	0.366	0.366	0.366	0.366	0.029	0.104	0.158	0.305	0.378
Transfer	**Clusters**	2	2	1	1	7	12	14	12	14
	Eval	0.902	0.829	0.788	0.718	0.461	0.259	0.246	0.288	0.331

	Silhouettes	0.249	0.249	0.296	0.296	-9.682	0.119	0.282	0.382	0.377
Import	**Clusters**	2	2	2	2	3	5	6	7	7
	Eval	0.862	0.843	0.765	0.687	0.472	0.368	0.392	0.376	0.39
	Silhouettes	0.227	0.219	0.219	0.219	0.078	0.214	0.474	0.403	0.403
StringConverter	**Clusters**	3	3	3	3	3	3	3	3	3
	Eval	0.942	0.885	0.827	0.77	0.712	0.655	0.597	0.54	0.482
	Silhouettes	0.425	0.425	0.425	0.425	0.425	0.425	0.425	0.425	0.425
RipCdDialog	**Clusters**	3	3	3	3	3	4	5	5	7
	Eval	0.927	0.855	0.783	0.711	0.638	0.495	0.465	0.418	0.409
	Silhouettes	0.277	0.277	0.277	0.277	0.277	0.307	0.307	0.314	0.362
DefaultDrawingView TransferHandle	**Clusters**	3	3	3	3	3	3	3	3	3
	Eval	0.922	0.845	0.768	0.69	0.613	0.536	0.458	0.381	0.304
	Silhouettes	0.227	0.227	0.227	0.227	0.227	0.227	0.227	0.227	0.227
MDIApplication	**Clusters**	5	5	5	5	5	5	5	6	5
	Eval	0.923	0.846	0.77	0.693	0.616	0.54	0.463	0.375	0.31
	Silhouettes	0.233	0.233	0.233	0.233	0.233	0.233	0.233	0.239	0.233

V. Result and Discussion

A. Result of the Experiment

The proposed algorithm was implemented in the prototype applications. Ten study cases of the Blob class are ready to use to ensure the new approach's final result. All of the classes were decomposed using a prototype application that was already updated using a new algorithm. Every result using the static and dynamic threshold decomposition is described in the following tables (Table II for the static and Table III for the dynamic threshold).

B. Compared to the Previous Approach

In the previous paper, two study cases were used, one of which is MDIApplication class. The result of decomposition using the previous algorithm on MDIApplication (using $a = 0.5$ and $b = 0.5$) is as described in Table IV and V.

In the case of the Silhouette value, using a new approach (after adding the validation process using the $Eval$ value), the result is shown as case number 10 (Tables II and III). There are increments of Silhouettes value of both static and dynamic threshold decomposition. The static threshold increased from 0.08 to 0.274, and the dynamic threshold increased from 0.15 to 0.233. Even though the dynamic threshold has a lower Silhouettes value, the dynamic threshold produces more clusters that match the purpose of single responsibility principles. More clusters are produced using a dynamic threshold.

The other result compared to the previous approach is the useless class. The problem emerged according to the cluster that only has one element (Table V), and the element is private (cluster number 2). The element name is scrollPane which has -0.27 of Silhouette. After updating the algorithm using the evaluation process ($Eval$ value), the result of decomposition is shown as follows (Table VI). The scrollPane element is moved to cluster number five, together with the other element. No clusters are considered unusable after updating the algorithm using the evaluation process.

C. Discussion

The previous section shows the experiment result after updating the algorithm using the evaluation process. Tables II and III show the detail of every combination of weight and express every case based on the cluster, $Eval$ value, and Silhouettes value. The result is different from one case to the other. For example, six cases using the dynamic threshold AHC produced a better value of Silhouettes than the static threshold AHC. Those cases are AudioFile, Interpreter, SVGOutputFormat, Import, and RipCDDialog. The rest of the cases are better using the static threshold AHC. The Silhouettes value is used as the consideration because the cluster requirement results in the high compactness of elements based on the similarity of syntax and semantics.

Higher Silhouettes also show the similarity of the cluster's context to produce the single responsibility class. With the use of the evaluation process, the result of Silhouettes can be increased by at least 40% of Silhouettes. This result shows that the evaluation process can place the elements more precisely by considering the value of the class usability and the Silhouettes. Most of the results show that the combination of weight (a and b) that can produce the best cluster is a higher portion than b.

TABLE IV. The Static Decomposition (Previous Approach)

Cluster	**Elements**	**Silhouettes Index**
1	parentFrame	-0.12
	MDIApplication	-0.03
	desktopPane	0.01
	Show	-0.41
	isSharingToolsAmongViews	-0.01
	Hide	-0.39
	serialVersionUID	-0.03
	scrollPane	0.03
	Prefs	0.00
2	createFileMenu	0.30
	Init	0.01
	getComponent	0.06
	createViewActionMap	0.31
	Configure	0.05
	createModelActionMap	0.15
	toolBarActions	-0.01
	createViewMenu	0.30
	updateViewTitle	0.32
	createHelpMenu	0.34
	createWindowMenu	0.30
	initLookAndFeel	0.11
	wrapDesktopPane	0.04
	createMenuBar	0.21
	createEditMenu	0.32
	Launch	0.05
Average Silhouettes		**0.08**

TABLE V. THE DYNAMIC DECOMPOSITION (PREVIOUS APPROACH)

Cluster	Elements	Silhouettes Index
1	isSharingToolsAmongViews	-0.12
	Prefs	-0.08
2	scrollPane	-0.27
3	parentFrame	0.00
	desktopPane	0.03
4	MDIApplication	0.27
	serialVersionUID	0.22
5	Show	-0.19
	Hide	-0.12
6	getComponent	-0.51
7	Launch	-0.62
8	createFileMenu	-0.84
	Init	-0.39
	initLookAndFeel	-0.49
	createMenuBar	-0.58
9	updateViewTitle	0.28
	Configure	0.08
10	createViewMenu	0.78
	createHelpMenu	0.89
	createWindowMenu	0.75
	createEditMenu	0.73
11	wrapDesktopPane	0.95
	toolBarActions	0.98
12	createViewActionMap	1.00
	createModelActionMap	1.00
Average Silhouettes		**0.15**

TABLE VI. RESULT OF DECOMPOSITION USING EVALUATION PROCESS ($a = 0.5; b = 0.5$)

Cluster	Elements	Modifier	Silhouettes
1	MDIApplication	publicMethod	0.083
	serialVersionUID	private	0.128
	isSharingToolsAmongViews	publicMethod	0.066
2	init	publicMethod	0.31
	initLookAndFeel	protectedMethod	0.149
3	updateViewTitle	protectedMethod	0.073
	configure	publicMethod	0.203
	launch	publicMethod	0.128
	show	publicMethod	0.087
	hide	publicMethod	0.053
4	createViewMenu	publicMethod	0.532
	createHelpMenu	publicMethod	0.59
	createWindowMenu	publicMethod	0.527
	createEditMenu	publicMethod	0.483
	createFileMenu	publicMethod	0.497
	createMenuBar	protectedMethod	0.326
	createViewActionMap	protectedMethod	0.572
	createModelActionMap	protectedMethod	0.282
	parentFrame	private	0.03
5	wrapDesktopPane	protectedMethod	0.331
	toolBarActions	private	0.159
	getComponent	publicMethod	0.059
	desktopPane	private	0.091
	scrollPane	private	0.035
	prefs	private	0.023
Average Silhouettes			**0.233**

TABLE VII. ELEMENT'S CHARACTER OF STUDY CASES

Class	A	B	C	D
AudioFile	39	9	30	Dynamic
JDBCBench	33	21	12	Static
Interpreter	65	20	45	Dynamic
SVGOutputFormat	61	9	52	Dynamic
Transfer	80	50	30	Dynamic
Import	30	13	17	Dynamic
StringConverter	16	1	15	Static
RipCdDialog	36	15	21	Dynamic
DefaultDrawingViewTransferHandle	15	2	13	Static
MDIApplication	25	6	19	Static

A: Total element; B: Attribute Element; C: Method Element; D: Approach

TABLE VIII. CORRELATION RESULT

No.	Pair Data	*p*-value
1.	Element - Approach	0.0134
2.	Attribute Element - Approach	0.1645
3.	Method Element - Approach	0.0247

In the specific number, $a \geq 0.7$ is suitable to produce a better cluster in both static and dynamic threshold AHC. Six cases are good using a dynamic threshold, and four cases using a static threshold. This result raises curiosity about whether the class decomposition uses static or dynamic.

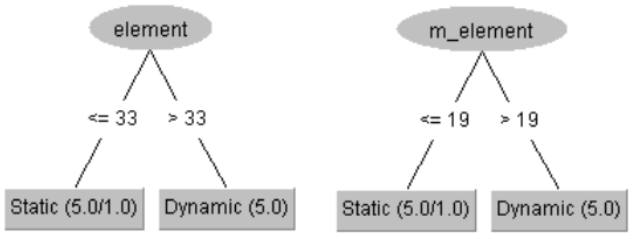


Fig. 5. The Tree Visualization of Tree-Based Classification Analysis

Based on the existing study cases in this research, every studied case is detailed into specific characteristics of class that relate to the element. For example, the number of total elements, the number of method elements, and the number of attribute elements are counted to find a correlation with the type of approach. Table VII shows the detail of the class based on the element.

The correlation between each character to the approach that is used on the class decomposition is counted using a statistical approach. There are three data pairs; the result is shown in Table VIII. Two pairs of data have significant differences. It is determined based on the result of the p-value of each pair. The total element (No. 1) and method element (No. 3) has a p-value lower than 0.05, and the attribute element (No. 2) is higher than 0.05. The total element and the number of method elements are related to the type of approach used in the class decomposition process.

The threshold number that can be used as the decision point for each total element and method element is also interesting. The data of element characteristics were analyzed using the tree-based classification method (Fig. 5). Furthermore, the total element and method element can indicate when the static or dynamic threshold should be used in the class decomposition process. The tree visualization shows the threshold number for each characteristic. The total element and method element have the threshold numbers 33 and 19. If the number of each characteristic is lower than the number, then the static threshold AHC is better and vice versa. This classification analysis result has an accuracy of about 80%.

The statistical and the threshold number analysis is only suitable for the current scope of the experiment. It needs more study cases to make the result acceptable to the larger scope of the experiment.

VI. Conclusion and Future Work

The class decomposition in the level of design is worth doing to support the concept of model-driven software engineering. The optimization of the design level Threshold-based Agglomerative Hierarchical Clustering (AHC) experiment has been done by adding an evaluation process. The evaluation process aims to move the specific element with negative Silhouettes value in every cluster to the other better cluster. The evaluation process is able to increase the average Silhouettes of the cluster compared to the previous approach. The increment of Silhouettes has averaged about up to 40%. The evaluation process is also able to solve the unusable cluster, as mentioned in the previous approach result.

This research experiment takes ten study cases from the Landfill smell dataset. All data consists of Blob smell. Most of the good result of decomposition is using the Dynamic Threshold AHC. Six study cases are good using the dynamic threshold, and four study cases are good using the static threshold. And the best result is produced by using the higher portion of Silhouettes (a) in the Eval formula.

The results were analyzed using a statistical approach to get more valuable information about the result. Three variables are related to the class: the number of total elements, method, and attribute elements. The total element and method element affect the use of the approach (static or dynamic threshold) to get a better result of decomposition. In the scope of the experiment, both the total element and method element have the threshold numbers 33 and 19. A smaller number than the threshold will use static, and a larger will use dynamic threshold.

The design-level class decomposition research is important to be continued in the future. The future plan aims to increase the optimization of the result of decomposition. In this experiment, the evaluation process is a separate process that is in sequence with the previous approach. Merging the algorithms became a consideration for future computational improvements. The increment of study cases number is worth increasing the algorithm's usability. Implementing the decomposition process at the source code level is worth doing in the future. The impact of changes in real implementation is important to study.

References

[1] M. Fowler et al., "Refactoring Improving the Design of Existing Code Second Edition," Second Ed. United State of America: Pearson Education - Wesley, 2019.

[2] M. Brambilla, J. Cabot, and M. Wimmer, "Model-driven software engineering in practice." Morgan & Claypool, 2012.

[3] A. Yamashita and L. Moonen, “Exploring the impact of inter-smell relations on software maintainability: An empirical study,” in Proceedings - International Conference on Software Engineering, 2013.

[4] F. Palomba, G. Bavota, M. Di Penta, F. Fasano, R. Oliveto, and A. De Lucia, “On the diffuseness and the impact on maintainability of code smells: a large scale empirical investigation,” Empir. Softw. Eng., vol. 23, no. 3, pp. 1188–1221, Jun. 2018.

[5] B. Priyambadha and T. Katayama, “Design Level Class Decomposition using the Threshold-based Hierarchical Agglomerative Clustering,” Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 3, pp. 57–64, 2022.

[6] B. Priyambadha and T. Katayama, “Tree-based keyword search algorithm over the visual paradigm’s class diagram XML to abstracting class information,” 2020 IEEE 9th Glob. Conf. Consum. Electron. GCCE 2020, pp. 280–284, 2020.

[7] B. Priyambadha, T. Katayama, Y. Kita, H. Yamaba, K. Aburada, and N. Okazaki, “Utilizing the similarity meaning of label in class cohesion calculation,” J. Robot. Netw. Artif. Life, vol. 7, no. 4, pp. 270–274, 2021.

[8] B. Priyambadha, T. Katayama, Y. Kita, H. Yamaba, K. Aburada, and N. Okazaki, “The Seven Information Features of Class for Blob and Feature Envy Smell Detection in a Class Diagram,” 2021 Int. Conf. Artif. Life Robot., pp. 348–351, 2021.

[9] K. Alkharabsheh, Y. Crespo, E. Manso, and J. A. Taboada, “Software Design Smell Detection: a systematic mapping study,” Softw. Qual. J., vol. 27, no. 3, pp. 1069–1148, 2019.

[10] B. K. Sidhu, K. Singh, and N. Sharma, “A Catalogue of Model Smells and Refactoring Operations for Object-Oriented Software,” Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018, pp. 313–319, 2018.

[11] B. Kaur Sidhu, “Model Smells In Uml Class Diagrams,” Int. J. Enhanc. Res. Manag. Comput. Appl., vol. 5, pp. 2319–7471, 2016.

[12] R. C. Martin, Clean Architecture: A Craftsman’s Guide to Software Structure and Design. 2017.

[13] Y. Wang, H. Yu, Z. Zhu, W. Zhang, and Y. Zhao, “Automatic Software Refactoring via Weighted Clustering in Method-Level Networks,” IEEE Trans. Softw. Eng., vol. 44, no. 3, pp. 202–236, 2018.

[14] G. Bavota, A. De Lucia, A. Marcus, and R. Oliveto, “A two-step technique for extract class refactoring,” ASE’10 - Proc. IEEE/ACM Int. Conf. Autom. Softw. Eng., pp. 151–154, 2010.

[15] G. Bavota, A. De Lucia, and R. Oliveto, “Identifying Extract Class refactoring opportunities using structural and semantic cohesion measures,” J. Syst. Softw., vol. 84, no. 3, pp. 397–414, 2011.

[16] G. Bavota, A. De Lucia, A. Marcus, and R. Oliveto, “Automating extract class refactoring: an improved method and its evaluation,” Empir. Softw. Eng., vol. 19, no. 6, pp. 1617–1664, 2014.

[17] M. Fokaefs, N. Tsantalis, A. Chatzigeorgiou, and J. Sander, “Decomposing object-oriented class modules using an agglomerative clustering technique,” IEEE Int. Conf. Softw. Maintenance, ICSM, pp. 93–101, 2009.

[18] M. Fokaefs, N. Tsantalis, E. Stroulia, and A. Chatzigeorgiou, “Identification and application of Extract Class refactorings in object-oriented systems,” J. Syst. Softw., vol. 85, no. 10, pp. 2241–2260, 2012.

[19] M. Hamdi, R. Pethe, A. S. Chetty, and D. K. Kim, “Threshold-driven class decomposition,” Proc. - Int. Comput. Softw. Appl. Conf., vol. 1, pp. 884–887, 2019.

[20] I. Bassey, N. Dladlu, and B. Ele, “Object-Oriented Code Metric-Based Refactoring Opportunities Identification Approaches: Analysis,” Proc. - 4th Int. Conf. Appl. Comput. Inf. Technol. 3rd Int. Conf. Comput. Sci. Appl. Informatics, 1st Int. Conf. Big Data, Cloud Comput. Data Sci., pp. 67–74, 2017.

[21] Á. Domingo, J. Echeverría, Ó. Pastor, and C. Cetina, “Evaluating the Benefits of Model-Driven Development,” Adv. Inf. Syst. Eng., no. June

2021, pp. 353–367, 2020.

[22] R. Dijkman, M. Dumas, B. van Dongen, R. Käärik, and J. Mendling, “Similarity of business process models: Metrics and evaluation,” Inf. Syst., vol. 36, no. 2, pp. 498–516, 2011.

[23] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, pp. 53–65, 1987.

[24] F. Palomba et al., “Landfill: An Open Dataset of Code Smells with Public Evaluation,” 2015 IEEE/ACM 12th Work. Conf. Min. Softw. Repos., pp. 482–485, 2015.

A Multi-Objective Optimization for Supply Chain Management using Artificial Intelligence (AI)

Mohamed Hassouna[1]
Management Information System Dept., Higher Technological Institute, HTI, 10th of Ramadan
Egypt

Ibrahim El-henawy[2]
Computer Science Dept Faculty of Computers and Information
Zagazig University
Zagazig, Egypt

Riham Haggag[3]
Business Information System Dept.
Faculty of Commerce and Business Administration, Helwan University
Helwan, Egypt

***Abstract*—Supply chain management seeks to solve the complex problems of transporting goods from the suppliers to the end customers. Improving the differentiation between different paths to reduce costs and time may require smart systems. This paper proposes two new algorithms for determining, with Multi-Objective Optimization, the least cost and the most appropriate path between two nodes. First: Ant colony optimization (ACO) algorithm, working alongside with Multi Objective Optimization (MOO), is adopted to determine the shortest path and time between two nodes to reach with the least cost. Multi-Objective intelligent Ant Colony (MOIAC) algorithm improves supply chain management to achieve the optimal and the most appropriate solutions. Second: Particle Swarm Optimization (PSO) algorithm, also working alongside MOO, is adopted to determine the least cost, time, and shortest path. Multi Optimization Intelligent Particle Swarm (MOIPS) algorithm improves supply chain management by determining the shortest path with the least cost. These two proposed algorithms seek the optimal solution by MOO using a JAVA Program. The experimental results show the excellence of the first algorithm in determining the optimal and the most appropriate path while getting throw risks inherent in transporting goods. It also demonstrates excellence in transporting goods in the shortest possible time and with the least cost. The second algorithm also shows excellence in transporting goods with the least possible cost via the shortest path and in the shortest time.**

Keywords*—Supply chain management; artificial intelligence; particle swarm optimization; ant colony optimization and multi-objective optimization

I. Introduction

Now-a-days, Supply Chain (SC) networks play a key role among suppliers [1] and end customers. Generally, SC networks involve variant agents such as suppliers, manufacturers, distributors, wholesalers, retailers, and customers [2], beside the interactions between them. SC is more complicated than traditional logistics as it is not limited to the transportation process among variant agents; rather, it has different phases and roles for different agents, such as what is supplied by suppliers [3-5], and what is ordered by customers. SC networks are sophisticated supplier-customer networks encompass agents, information, techniques, activities, [6], and resources. SC networks consist of: suppliers, manufacturing or production factories, stores, distribution centers [7], and customers. This network aims to achieve optimal resource choice to reduce cost and time [8]. SC networks are the main structure of the operations and the interactions among those agents, from the preliminary strategic level [9], to the final operational one. A good practiced SCM is a competitive advantage for organizations working in the field of investment and raising capital. Organizations have variant options in managing such interactions in SC (supplying with goods, assessing products, offering end products to customers) [10-13], according to time, cost, and profit. The problem is that SCM is responsible for a huge number of processes and operations such as production and procurement planning, choosing the optimal product, customer orientation, marketing, distributing products [14], and sales among others. SCM has to balance the SC and each organization's different objectives; some objectives may contradict other objectives in the same organization. So, there must be an appropriate method to coordinate between such objectives taking into consideration that the SC has variant agents in variant phases (i.e. the supplier, the distributor, the seller, and the customer). Suppliers and end customers may have different locations, a thing that may increase the cost of transporting goods in different paths [15-17], and among different nodes to reach the end customer. To achieve this balance among different objectives, companies must consider comparing and differentiating between different timings & time limits and between the added costs for the goods to determine the appropriate path, cost, [18], and timing. Generally, it is clearly noted from previous relevant works and papers that SCM has many dimensions that need to be studied simultaneously to achieve the least cost and [19], the shortest time. In this paper, however, we not only focus on the least cost and the shortest time, but we also try to determine the optimal and the best path alongside with the highest profit while preserving the quality, and improving it if possible. Moreover, this paper focuses on reducing the cost while giving attention to possible risks that may occur in the transportation process. So, we must be precise and careful in improving SCM using the two new algorithms to reach the best possible results, then comparing them to those of other algorithms. Artificial intelligence techniques can help organization improve their objectives) [20]. (the cost - the time limit - the optimal path.In this Work, we use several objectives integrated with AI techniques (i.e. PSO and ACO algorithms)[21-24]. Problem description of SCM covers a wide range of subject. Users, distance, marketing, distribution, least cost path, production and procurement in companies work independently and in parallel in the supply chain. Although each of these companies

has its own objectives and often these objectives are in contradiction with another, so there needs to be a method to achieve these different objectives. We study the problem of the least cost and the shortest time path to improve the transportation process. We propose two algorithms for determining the least cost path: the first algorithm (MOIAC) determines with ZIPF random distributions the shortest and optimal path between nodes to reach to the end customer, while the second algorithm (MOIPS) reduces time and cost in the process of transporting goods from suppliers or producers to end customers. The Main Contribution of the study can be highlighted as follows:

- Solves the problem of determining the optimal path with the least cost to reach the end customer.
- Uses (MOIAC) algorithm to reduce the cost and shrink the distance by choosing the shortest path to the end customer, seeking a balance between the nodes.
- Uses Zipf distributions along with ACO to create random distributions to determine the optimal path to transport goods to the end customer in an appropriate time.
- Uses (MOIPS) algorithm to determine the optimal path to the end customer in the shortest time.
- Apply and test the two proposed algorithms using a JAVA program to verify their superiority over other algorithms.

The rest of this paper is organized as follows. Section I is the introduction. Section II overviews the relevant previous works that addresses the SCs. Section III focuses on the main structure proposed for transporting goods from suppliers to end customers. Section IV focuses on the proposed system; (MOIPS) and (MOIAC) algorithms. Section V introduces the experimental results and compares it with results of other algorithms. Section VI shows the conclusions and recommendations for further research.

II. Related Work

E. Mastrocinque et al. proposes a technique of improving SC using the Bees algorithm with MO to reduce the cost and the time consumed. It also uses the Pareto system to determine the optimal solutions to the problem of cost and time to improve the SC. It proposes new weights in applying the algorithm and compares the proposed system with other algorithms. The results show that the proposed algorithm exceeds other algorithms in reducing the cost and time. This work recommends complicating the problem and improving the Bees algorithm by integrating other objectives in further work [25].

R. Ehtesham et al. improves the SC by integrating other environmental and economic dimensions to the MO. The main goal of this work is to achieve the highest margin of profit by transporting the largest amount of goods, while reducing environmental pollution. This problem has been solved using two algorithms with Multi-Objective Optimization to select the suppliers and to improve the SC. These proposed algorithms have been applied to Mega Motor Company to reduce the cost and time. The results show that the proposed algorithm exceeds other algorithms in reducing environment pollution. This work recommends improving the algorithm using Meta-heuristic and addressing cost and time simultaneously [26].

H. Banerjee et al. proposes a new technique that is using Pareto Optimization in the cases of the uncertainty of the preliminary assumptions. It uses Pareto Optimization with a Genetic algorithm and Mixed-Integer Linear Programming (MILP). It shows some scenarios of avoiding risks in SC systems that are affected directly by the customer's requirements. The work also improves the process of selecting the nearest suppliers to the customers to reduce the total cost and to avoid risks. This methodology proves that the experimental results are better than those of other algorithms in cases of uncertainty. The work recommends using different algorithms to improve the methodology used in the cases of uncertainty [27].

L. Martínez et al. this work proposes the technique of Meta-heuristic using Water Drop with MO to reduce the cost and the time. It depends on Pareto optimization to determine the number of optimal solutions simultaneously. The results show that the proposed algorithm exceeds other algorithms in reducing the cost and time. This work recommends improving the algorithm using the distances between nodes to determine the optimal and the shortest path [28].

S. Gupta et al. proposes a method of an optimal allocation of suppliers and resources with specific products with the help of decision makers. The work divides the decision makers into two groups: the first group is responsible for the goods transported to distributors, and the second group determines the amounts reasonably. The first group is concerned with transporting goods with least cost, while the second group is concerned with reducing delivery time also with least cost. This paper uses Fuzzy with MO to address Conflicting objectives, reaching a compromise in the process of transportation. The results show that the proposed algorithm exceeds other algorithms in achieving the optimal amounts of products in the process of making a decision. This work recommends using Meta-heuristic with Pareto optimization [29].

R. Sun et al. describes the application of Ant Colony with MO in SCM. It addresses a number of objectives such as cost, time, customer service, and flexibility with the goal of improving the SC. The work also introduces MO system to solve some problem to improve the SC. It recommends improving the algorithm and using other algorithms [30].

P. Phuc et al. focuses on the problem of directing the vehicles for logistic services. While delivering a product to the customer, the vehicle has to pass over all the nodes inherent in the network to reach every customer in their lists. The main objective of this work is to reduce the cost of traveling from one customer to another, considering that not all vehicles are similar. ACO has been used to direct vehicles and detect each vehicle's arrival time. The work recommends analyzing more optimal results by integrating MO and using AI to reach the shortest path, considering time and traffic [31].

Y. Wenfang et al. has designed a new strategy to manage the inventory of the SC, manage the marketing process, and improve companies' response speed. It also improves more than one methodology of ACO algorithm with Fuzzy. This work positively influences the efficiency of the organizations' ability to manage inventory in SC. The work recommends using AI to manage inventory to improve SC [32].

X. Zhang et al. developed ACO algorithm with MO using two different colonies to reduce the cost of the goods in the SC. The work also develops a method to determine priorities and weights, detecting the path of transporting goods and the optimal cost. The results show that the proposed system exceeds other algorithms on a large scale in smart cities. Therefore, this work recommends reducing resource consumption to the minimum, and improving the system with other algorithms that can be applied on a larger scale with addressing objectives such as cost, time, and optimal path to transport goods [33].

A. Discussion and Related Works

It is clearly noted from previous studies and articles that is relevant to this field that SCM has many dimensions that has been largely studied to achieve the least cost and the shortest time. This paper does not only focus on reducing the cost and time, but it also tries to determine the optimal and the best path taking into consideration the highest margin of profit and preserving the quality of the product and improving it without negatively affecting the customer or environment. The paper focuses on reducing the cost while giving attention to possible risks that may occur in the transportation process among nodes as presented in Table I. So, we have to be precise and careful in improving SCM using the two proposed algorithms (PSO & ACO) to reach optimal results.

TABLE I. RELATED AND THE PROPOSED WORK COMPARISON

Strategy	Year	AI Techniques	Distance	Least Cost Path
Bees Algorithm	2013	√	√	
Intelligent Water Drop(IWD)	2014	√	√	
Ant Colony Algorithm and Fuzzy Model	2019	√	√	
Mixed-Integer Linear Programming (MILP)	2020	√	√	
multi-objective particle swarm optimization (MOPSO)	2020	√	√	
mult iobjective supply chain configuration (MOSCC)	2021	√	√	
My Propsed	2022	√	√	√

III. PROPOSED ARCHITECTURE FOR SUPPLY CHAIN DESIGN CASE STUDY USING THE SWARM INTELLIGENT WITH MULTI-OBJECTIVE

Presented in Fig. 1, this section describes the proposed structure of the smart system of a SC from the supplier to the customer, where agents are referred to by nodes on the network [11]. In our model, we use a Heterogeneous system with AI techniques and MO to determine the optimal path in transporting goods to reach the end customer, considering time and cost problems. The proposed system consists of (i) suppliers (the first node, from which goods are transported via different types of vehicles and different paths to reach the next node), (ii) distributors (the second node, the wholesaler who receives goods from suppliers, classify them, then transport them to the next node), (iii) retailers (the third node, who finally hand the goods over to the next node), and (iv) end customers (the final node). This is clearly shown in the figure.

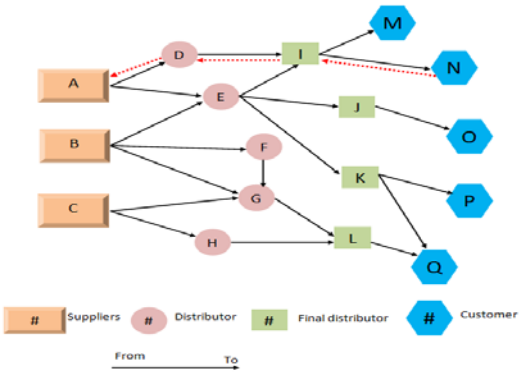


Fig. 1. Proposed Model Architecture for Supply Chain Management Systems.

Determining the location of the optimal supplier to the customers depends on the latter's needs. Choosing the optimal and the shortest path is accomplished using an AI and a number of mathematical equations concerning time, cost, and distance. When goods are to be transported from suppliers to customers, the proposed algorithm determines the optimal path. The proposed system is divided into four parts: (i) using AI techniques to choose the optimal path, (ii) using an improved ACO algorithm with MO, (iii) using a PSO algorithm with MO to improve the system, and (iv) employing the equations of cost, time, and distance among nodes to reach the destination with the least cost. Finally, the proposed system is applied using a JAVA program.

A. Multi-Objective Optimization in Supply Chain

MOO in the SC is improved using AI techniques to determine the optimal and shortest path among nodes. To accomplish such objectives in the SC, cost and transportation systems have to be improved. That is why we integrate the distance equation among nodes to determine the optimal and shortest paths among suppliers and customers [33].

1) Cost: Costs among different nodes are calculated to reach the optimal cost. Costs of transporting goods from suppliers and customers must be low for the variant means of transportation [34]. We need to consider that the system is a Heterogeneous system.

$$\mathbf{Cost}(\boldsymbol{c_i}) = \sum_{i=1}^{n}\left(\boldsymbol{cost}(\boldsymbol{c_{ij}}) * \boldsymbol{demand}(\boldsymbol{d_{ij}})\right) \quad (1)$$

$$\boldsymbol{c_{ij}} \in \{0,1\}, ij = 1,2,\dots,\text{n } Node$$

2) Distance: The minimum distance among nodes is calculated to determine the optimal and the shortest path among nodes in the system to ensure that goods are handed over to the end customers through the path with the least cost and time.

$$\boldsymbol{Min}\ \sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{d_{ij}x_{ij}} \quad (2)$$

S.T:

$$\sum_{\mathbf{i=1}}^{\mathbf{n}}\boldsymbol{n_i x_{i\geq k, x_i\in\{0,1\}\ (1\leq i\leq n)}} \quad (3)$$

3) Time : Time needed to move among different nodes is calculated to reach the optimal time. $LTime_i$ is the time among the nodes, while $_i$ is the time of the node. The goal is to shrink the time consumed when transporting goods among nodes.

$$\mathbf{Demand_i} = \sum_{\mathbf{j\in\{downNodei\}}}\mathbf{Demand\ j, i = 1,2,\dots\dots,nNode} \quad (4)$$

$$\mathbf{LTime_i = Time_{i,xi} + MaxUPLT_i}\ \mathbf{, i = 1,2,\dots\dots,nNode} \quad (5)$$

$$\mathbf{MaxUPLT_i = Max_{j\in\ UpNodei}}\left\{\mathbf{LTime_j}\right\}\mathbf{, i = 1,2,\dots\dots,nNode} \quad (6)$$

Equations (4) and (6) show these functions: the time needed for the SCN to accomplish the work is referred to by T, the total number of the nodes in the network is nNode, the demand quantity of node is $Demand_i$, while the time of the $node_i$ is $LTime_i$. Equation (5) shows the decision vector x with node, where the number of options available for node $_i$ is $Option_i$, and the chosen option of the corresponding node is represented by different values of dimensions. Equation (6) calculates $Demand_i$, where the set of down nodes of node $_i$ is $downNode_i$, and it is previously determined for each customer. Equation (5) calculates $LTime_i$, where the time needed of the node $_i$ to accomplish option $_{xi}$ is $Time_{i, xi}$, the maximum LTime of $upNode_i$, is $maxUpLT_i$ (the set of up nodes calculated as shown in equation (6). The typology of SCN determines the upNode and the downNode.

4) ZIPF Distribution: Zipf distributions create random distributions of goods among nodes. Goods are distributed among nodes according to the different tasks of the suppliers and the end customers [35].

$$\boldsymbol{p(f_i)} = \frac{\mathbf{1}}{\boldsymbol{i^a}} \quad (7)$$

Where i = 1, 2, …, n; and α is a factor of goods distribution, where $0 \leq \alpha < 1$.

IV. Proposed PSO and ACO-based Algorithm for the Supply Chain

A. Multi Objective with Particle Swarm Optimization

The process updates the particle velocity, position and inertia weight is presented in Table II using Eq. (8), Eq. (9) and Eq. (10) as follows [33-36]. We update the velocities for every particle as follows:

$$V_{i,j}^{k+1} = W.V_{i,j}^{k} + C_1R_1\left(pbest_{i,j}^{k} - X_{i,j}^{k}\right) + C_2R_2\left(gbest_{i,j}^{k} - X_{i.j}^{k}\right) \quad (8)$$

Where

$V_{i,j}^{k+1}$ Refers to the new velocity of a particle

$V_{i,j}^{k}$ Refers to current velocity

C_1, C_2 positive constants acceleration parameters

$pbest_{i,j}^{k}$ personal best position particle

$X_{i,j}^{k}$ position of i^{th} particle in j^{th} swarm

R_1, R_2 two random variables in the range [0,1]

$gbest_{i,j}^{k}$ global best position particle

$$\boldsymbol{X_{i,j}^{k+1} = X_{i,j}^{k} + V_{i,j}^{k+1}} \quad (9)$$

Where

$X_{i,j}^{k+1}$ new position of particle

k iteration population

$i \in 1,2,3,\dots,m$ m is the number of members in an iteration

$j \in 1,2,3,\dots,d$ d is the size of the swarm

$$\boldsymbol{w = w_{max} - \frac{w_{max}-w_{min}}{iter_{max}} * iter} \quad (10)$$

Where

w inertia weight

w_{max} initial value of inertia weight

w_{min} final value of inertia weight

$iter_{max}$ maximum number of iterations

$iter$ current iteration number

Accomplishing the mission of reaching the end nodes, the proposed algorithm is proven to choose the optimal nodes to reach the destination by testing the appropriateness of each node according to the agents in the network. The algorithm uses MOO to determine paths with the shortest distance and the least cost and time. The steps of the MOIPS Algorithm are shown in Fig. 2 and (Algorithm 1):

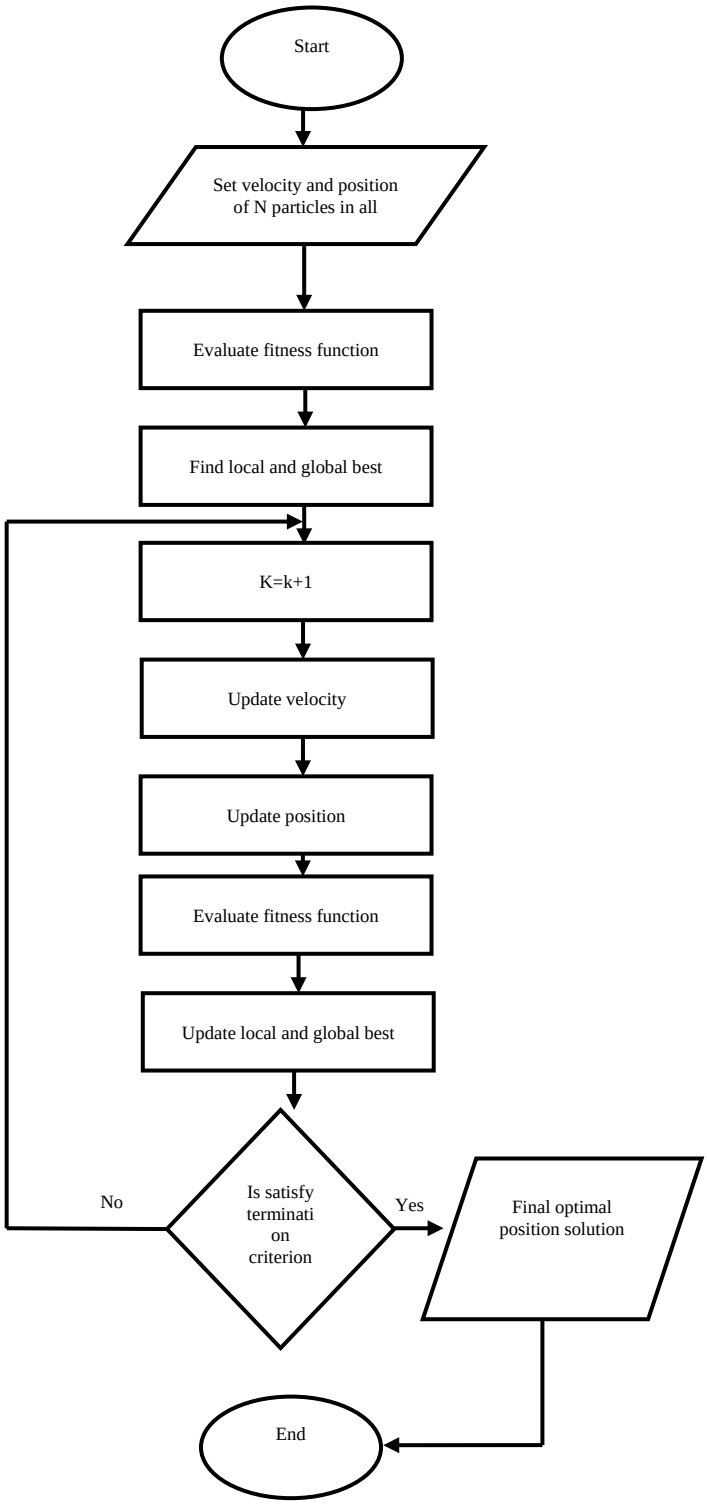


Fig. 2. Proposed Flowchart MOIPS.

Algorithm 1. The Proposed MOIPS

Input: Size α of Population
Number of Iterations
Node
Cost and Time
Distance
Output: Selected $P_{position}$ ← (Optimal $_{Best\ node}$, Optimal $_{Total\ Execution\ Time}$ and Optimal $_{Costs}$)
Initialization:
Define Values of parameters, Size of Pop, Num of Iterations and Num of Particles
Initialize set values of particle swarm (Num of Iteration)
Initialize availability and unavailability probabilities
Initialize best node according to costs and time
Repeat
Count I = 0
For j=1 to α do
For each goods in node do
Calculate fitness function
Update velocity
Update position
$P_{velocity}$ ←Random velocity()
$P_{position}$ ←Random position()
$P_{best} \leftarrow P_{position}$
If $\alpha \leq 0$ then
Exploitation
Else
Exploration
Select best node
End if
Calculate the distance of the node
Calculate the time of the node
Calculate the costs of the node
End for
End for
Until maximum number of iterations is reached, or access solution optimal
Return the optimal best node solution

TABLE II. PSO PARAMETERS

No.	Parameters	Values
1	*Number of particles*	100
2	*C1*	2
3	*C2*	2
4	*R1*	[0 - 1]
5	*R2*	[0 – 1]
6	w_{max}	0.9
7	w_{min}	0.4
8	*Number of iteration*	1000
9	*W*	1
10	*Population (swarm size k)*	50

B. Multi-objective with Ant Colony Optimization

In this section, the MOIAC algorithm is discussed. This algorithm determines the least cost path between nodes depending on time, cost, and optimal distance in order for the supplier to reach the end customer. MOIAC algorithm is applied to choose between one path or another to reach the optimal choice according to the needs of the end customers from the suppliers. MOIAC algorithm is of great benefit in reaching the shortest path with the least cost and time. Both ACO and PSO are types of swarm intelligence. The task of determining the optimal path with the least cost is NP-hard; however, it can be more useful in solving complicated problems than traditional methods as presented in Table IV [9-12]. The positions of the pheromones are calculated on different paths using FF, while moving from one node to another is calculated according to the following:

$$p_{ij} = \begin{cases} \frac{[\tau_{ij}]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{s\in k}[\tau_{is}]^{\alpha}[\eta_{ij}]^{\beta}} \\ 0 \; otherwise \;\; if \; j \in k \end{cases} \quad (11)$$

The calculation of the next node that is selected by Eq as follows:

$$\boldsymbol{i} = \begin{cases} \boldsymbol{arg\,max_{s\in k}} \{[\boldsymbol{\tau_{ij}}]^{\boldsymbol{\alpha}}[\boldsymbol{\eta_{ij}}]^{\boldsymbol{\beta}}\}, \boldsymbol{if\ q \le q_0}, \boldsymbol{if\ q > q_0} \\ \boldsymbol{j} \end{cases} \quad (12)$$

The calculation of the detection array of the ant proceeds according Eq. (13):

$$\boldsymbol{\eta_{ij}} = \frac{\mathbf{1}}{\boldsymbol{d_{ij}}} \quad (13)$$

The pheromone values on the routes are updated after every repetition. When ants reach the end of their travel path, the pheromone value is a positive constant. The updated local pheromone value can be calculated by Eq. (14) as follows.

$$\boldsymbol{\tau_{ij}} = (\mathbf{1} - \mathbf{p})\boldsymbol{\tau_{ij}} + \boldsymbol{p_{\tau 0}}, \forall(\mathbf{i},\mathbf{j}) \in \boldsymbol{t_k}, \mathbf{where}\ (\mathbf{0} < p \le 1) \quad (14)$$

After evaporation, every ant adds pheromones to the routes according to the set method, and the updated global pheromone value is calculated by Eq. (15) as follows:

$$\boldsymbol{\tau_{ij}} = (\mathbf{1} - \boldsymbol{p}) + \boldsymbol{p}.\sum_{\boldsymbol{k=1}}^{\boldsymbol{m}} \Delta_{\boldsymbol{\tau ij}} \quad (15)$$

$\Delta_{\tau ij}$: is the amount of pheromone added by ant k on their route. It can be represented Eq. (16) as follows:

$$\Delta_{\boldsymbol{\tau ij}} = \begin{cases} \frac{\mathbf{1}}{\boldsymbol{c^k}} \; \boldsymbol{if}\ \forall(\boldsymbol{i},\boldsymbol{j}) \in \; \boldsymbol{t^k} \\ \mathbf{0}\ \boldsymbol{otherwise} \end{cases} \quad (16)$$

$$\boldsymbol{p(f_i)} = \frac{\mathbf{1}}{\boldsymbol{i^a}} \quad (17)$$

Where i = 1, 2, …, n; and α is a factor determining the data access distribution, where $0 \le \alpha < 1$. As mentioned in Table III, notation of ant colony optimization.

TABLE III. NOTATION OF ANT COLONY OPTIMIZATION (ACO)

α	Pheromone trial control parameter τ_{ij}
β	Pheromone trial control parameter η_{ij}
S	Probability for slave ant $\forall s \in \{1,2,3,\dots,k\}$
K	Ants
Q	Random variable uniformly distributed in [0,1]
J	Random variable selected to the probability distribution with $(\alpha = 1)$
I	Currently node i
J	Choose go to node j
P	Evaporation rate
M	Number of ants
τ_{ij}	Pheromone density or node j (amount of pheromone between i & j)
η_{ij}	Heuristic information (importance between nodes i & j) $= 1/d_{ij}$
p_{ij}	Probability ant transits. Currently node i to node j
$p_{\tau 0}$	Initial pheromone
$\Delta_{\tau ij}$	Amount of pheromone that an ant adds to the path it has visited
c^k	The length of tour t^kWhich was built by the Ant k
t^k	Total lengths of path
t_k	Tour by ant k iteration
d_{ij}	Distance between two nodes

TABLE IV. ACO PARAMETERS

No.	Parameters	Values
1	α	1
2	β	2
3	P	0.3
4	Q	1
5	m	110
6	t_k	800
7	$p_{\tau 0}$	0.8

The algorithm determines the optimal nodes using Zipf and calculating the fitness function for each node,

ZIPF distributions are applied to create distributions to reach the optimal nodes and paths. ZIPF is a random distribution that aims to determine the optimal and shortest paths between the supplier and the customer. The function is as follows: The steps of MOIAC Algorithm that aims at improving its distributions are shown in Fig. 3 and (Algorithm 2):

Algorithm 2. The Proposed MOIAC

```
Input: Number of Ants
       Number of Iterations
       nodes
       Zipf Distribution
       Min Distance between nodes
Output: Selected Optimally Best distance (Optimal Best node
       Optimal Total Execution Time and Optimal Costs)
   Initialization:
       Define Values of parameters, Num of Iterations and Num
of Ants
       Initialize distance between nodes
       Initialize costs of nodes
       Initialize time of nodes
   Repeat
     For I=1 to (Num of ants)
         Step = step + 1
         Set all ant distribution in node
       For each node in current system
         Calculate desirability of the movement
         Calculate probability of the movement
          If q ≤ q0 then
             Exploitation
          Else
             Exploration
          End if
       End for
       For each dimension do
         Calculate fitness function
         Update local pheromone
         Update global pheromone
         Set local pheromone update
         Set global pheromone update
         Set determine distance in nodes
         Until all nodes are selected
               If the least cost path is long
               Then
                Apply the global update rule
               Else if
                Apply this path
               End if
         End for
     End for
   Until max number of iterations is reached or access solution is
found
   Return the optimally best node
```

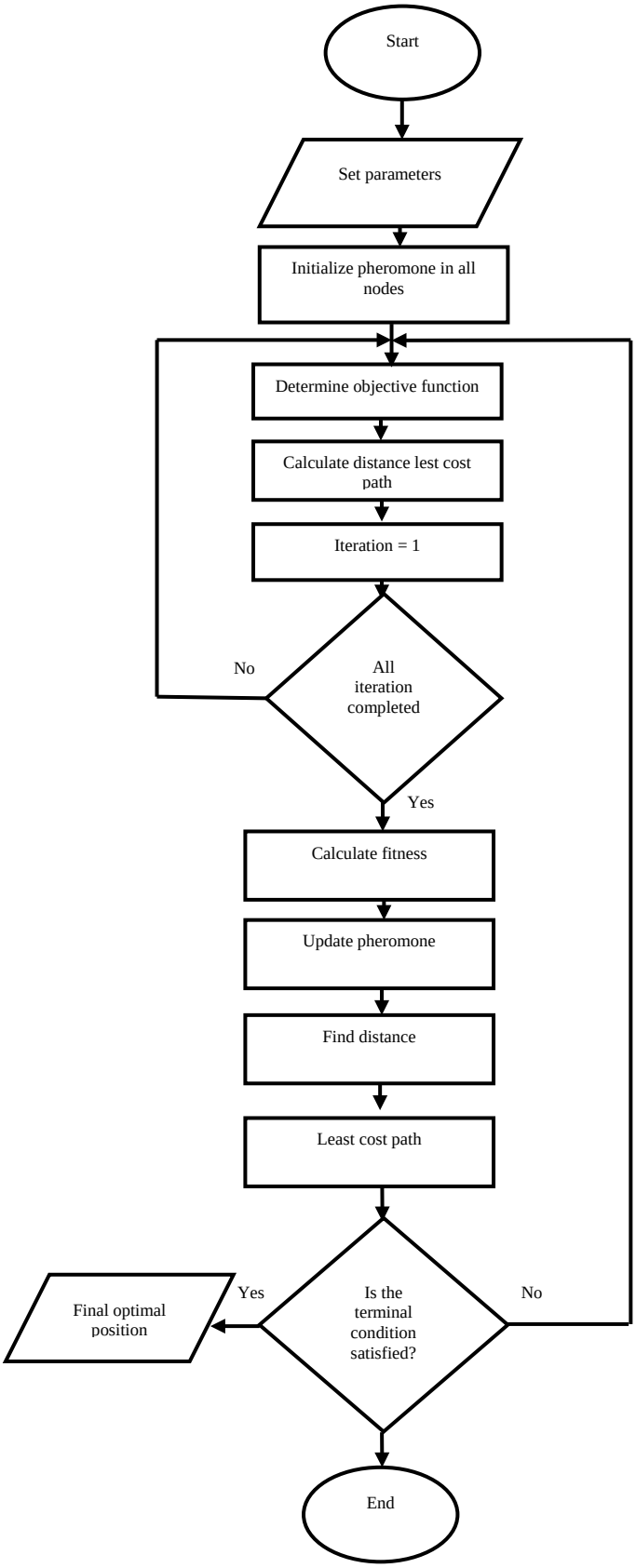


Fig. 3. Proposed Flowchart MOIAC.

V. Results and Discussion

This section discusses the experimental results of the model of determining the least cost path to reach the optimal and the most appropriate path using the proposed algorithms MOIPS and MOIAC in addition to Zipf random distributions. These algorithms are applied on a JAVA program. A comparison between these proposed algorithms and other algorithms has been held on the grounds of the time of their application, the cost, the time consumed, the high availability, determining the optimal and the most appropriate path, and the efficiency of the proposed system. The experiments have been carried out using a JAVA program that provides several classes to simulate and model the proposed system; we improve variant classes for the proposed system.

Fig. 4. Demonstrate the Cost for each Order.

In Fig. 4 compares between three algorithms on the grounds of the transportation rate and cost among suppliers and end customers. The algorithm determines the optimal and the shortest path according to MO with ACO and PSO. The experimental results show that MOIAC algorithm achieves the least cost when compared to MOIPS and MPACA algorithms. It is also the quickest in moving among nodes through the optimal paths. The results also show that the proposed algorithms have proven their excellence over the other algorithms according to the rate of goods transportation to the end customer.

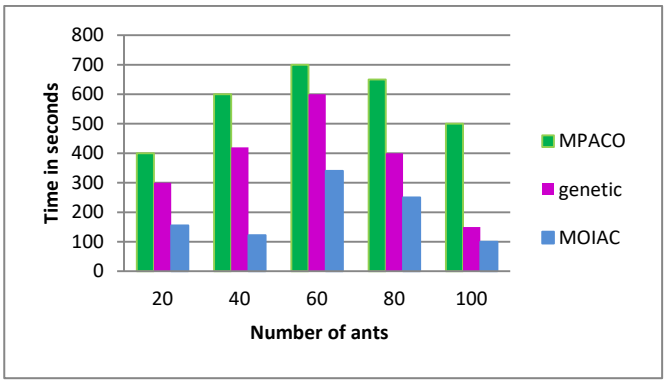


Fig. 5. Task Number of Ants.

Fig 5 shows that the proposed algorithm executes its missions in a lesser time when compared to the other algorithms. It also exceeds the other algorithms' performance when variant numbers of ants and scenarios are addressed. The results show that the proposed algorithm executes the scenario of 100 ants in a lesser time when compared to MPACS and Genetics algorithms.

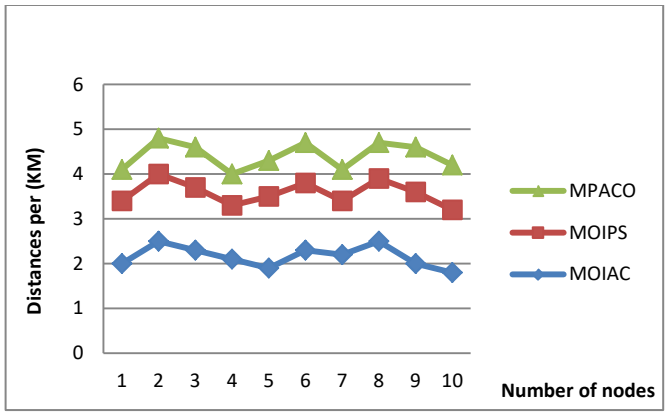


Fig. 6. Distances between the Number of Nodes.

Fig. 6 shows that the relation among the different algorithms reduces the time consumed and the cost of the transportation process done between the suppliers and the end customers using MOPSO and MOACO. The algorithm also considers determining the path with the least cost to reach the end customer. The experimental results show that the proposed algorithm surpasses the other algorithms.

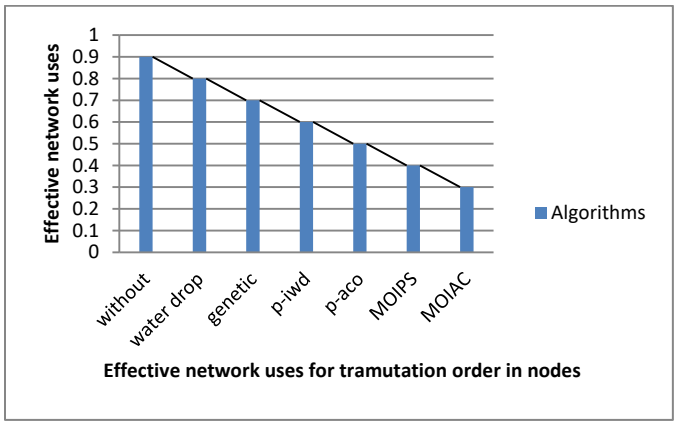


Fig. 7. Effective Network uses for Transmutation Order in Nodes.

Fig. 7 shows an effective version of the network and the percentages of goods crossing the nodes in the range of 0.1 to 1; the system detects the arrival time, the repetition frequency, and the response time among the nodes of the system. The improved bandwidth proves to be more effective with the proposed algorithm; it reaches 0.3 while it reaches 0.9 in the other algorithms. MOIAC algorithm surpasses the other algorithms on the grounds of efficiency, cost, and time.

Fig. 8 shows the determination of the optimal, shortest, and the least cost path among nodes, which positively affects the transportation process among the suppliers and the end consumers. When goods are ordered, the algorithm chooses the optimal, the shortest, and the least cost path among nodes on the proposed system. The proposed algorithm surpasses the other algorithms in the process of determining the optimal, the shortest, and the least-cost path. It is noteworthy that we have

tested the two proposed algorithms MOIAC and MOIPS, and the first surpasses the latter.

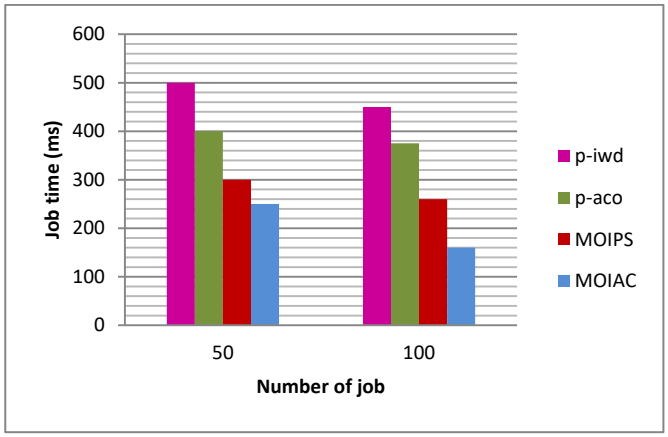


Fig. 8. Distances between the Number of Nodes.

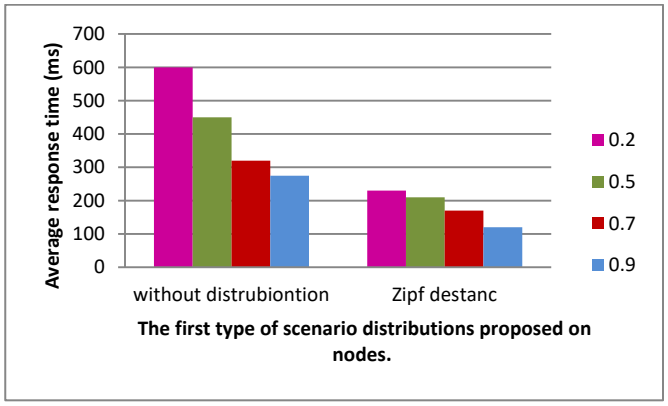


Fig. 9. Average Response Time of Zipf.

Fig. 9 shows the use of MOIAC algorithm with ZIPF distributions to determine the optimal and the most appropriate path to transport goods from the suppliers to the end customers. In determining the optimal path, variant distribution has ranged from 0.1 to 0.9. The experimental results prove the proposed algorithm's excellence in achieving optimal results in creating variant ZIPF distributions.

VI. Conclusion and Future Work

In this paper, we propose two algorithms to solve the problem of determining the optimal path between nodes, improving time-consumption and reducing cost simultaneously. The two proposed algorithms are used with Multi-Objective Optimization to improve the quality of transportation between the supplier and the end customers. MOIAC algorithm is designed to determine the optimal paths (the shortest and the least cost paths) among nodes. ZIPF distributions is integrated to create random distributions to reach the optimal nodes in each process. MOIPS algorithm is also designed to determine the optimal paths while reducing the time consumed in transporting goods from the suppliers to the end customers, and improving the transportation process following the least cost path. The proposed system has been tested on JAVA Program and has been also compared with other algorithms such as Water drop , genetic and bee.

Being integrated with Multi-Objective Optimization in the field of transportation and tested by AI techniques, the simulation results show the efficiency of the proposed algorithms. Many other objectives can be addressed in further works, such as improving means of transportation and reducing resource consumption using the least-cost paths. We also propose addressing other objectives, such as improving the cost, reducing the time consumed between the supplier and the end customer, speeding up the transportation process, and reducing risks. The two proposed algorithms are applicable with other objectives in the field of goods transportation.

References

[1] P. Fiala, “Information sharing in supply chains,” Omega, vol. 33, no. 5, pp. 419-423, Oct. 2005. doi:10.1016/j.omega.2004.07.006.

[2] F. Alawneh and G. Zhang, “Dual-channel warehouse and inventory management with stochastic demand,” Transport. Res. E-Log., vol. 112, pp. 81-106, Apri. 2018. doi:10.1016/j.tre.2017.12.012.

[3] M. Varsei and S. Polyakovskiy, “Sustainable supply chain network design: A case of the wine industry in Australia,” Omega, vol. 66, pp. 236-247, Jan. 2017. doi:10.1016/j.omega.2015.11.009.

[4] M. C. Chen, Y. H. Hsiao, and H. Y. Huang, “Semiconductor supply chain planning with decisions of decoupling point and VMI scenario,” IEEE Trans. Syst., Man, Cybern., Syst., vol. 47, no. 5, pp. 856–868, May 2017. doi:10.1109/tsmc.2016.2521740.

[5] Scheibe, P.Kevin., J. Blackhurst. ”Supply chain disruption propagation: a systemic risk and normal accident theory perspective.” International Journal of Production Research 56.1-2: 43-59. 2018. doi:10.1080/00207543.2017.1355123.

[6] K. Govindan, M. Fattahi, and E. Keyvanshokooh, “Supply chain network design under uncertainty: A comprehensive review and future research directions,” Eur. J. Oper. Res., vol. 263, no. 1, pp. 108-141, Nov. 2017. doi:10.1016/j.ejor.2017.04.009.

[7] Z. Hong, W. Dai, H. Luh, and C. Yang, “Optimal configuration of a green product supply chain with guaranteed service time and emission constraints,” Eur. J. Oper. Res., vol. 266, no. 2, pp. 663-677, Apr. 2018. doi:10.1016/j.ejor.2017.09.046.

[8] L. A. Moncayo–Martínez and Gustavo Recio, “Bi-criterion optimisation for configuring an assembly supply chain using Pareto ant colony meta-heuristic,” J. Manuf. Syst., vol. 33, no. 1, pp. 188-195, Jan. 2014. doi:10.1016/j.jmsy.2013.12.003.

[9] X. Zhang, et al., “Cooperative coevolutionary bare-bones particle swarm optimization with function independent decomposition for large-scale supply chain network design with uncertainties,” IEEE Trans. Cybern., vol. 50, no. 10, pp. 4454-4468, Oct. 2020. doi:10.1109/tcyb.2019.2937565.

[10] G. Zhang, J. Shi, S. S. Chaudhry, and X. Li, “Multi-period multi-product acquisition planning with uncertain demands and supplier quantity discounts,” Transport. Res. E-Log., vol. 132, pp. 117-140, Dec. 2019. doi:10.1016/j.tre.2019.11.005.

[11] Q. Long, X. Tao, Y. Shi and S. Zhang, “Evolutionary game analysis among three green sensitive parties in green supply chains,” IEEE Trans. Evol. Comput., vol. 25, no. 3, pp. 508-523, Jun. 2021. doi:10.1109/tevc.2021.3052173.

[12] G. Soni, V. Jain, F. T. Chan, B. Niu, and S. Prakash, “Swarm intelligence approaches in supply chain management: potentials, challenges and future research directions,” Int. J. Supply Chain Manage., vol. 24, no. 1, pp. 107-123, Jan. 2019. doi:10.1108/scm-02-2018-0070.

[13] L. A. Moncayo–Martínez and E. Mastrocinque, “A multi-objective intelligent water drop algorithm to minimise cost of goods sold and time to market in logistics networks,” Expert Syst. Appl., vol. 64, pp. 455-466, Dec. 2016. doi:10.1016/j.eswa.2016.08.003.

[14] A. Diabat and A. Jebali, “Multi-product and multi-period closed loop supply chain network design under take-back legislation,” Int. J. Prod. Econ., vol. 231, 107879, Jan. 2021. doi:10.1016/j.ijpe.2020.107879.

[15] H. Zhao, Z. G. Chen, Z. H. Zhan, S. Kwang, and J. Zhang, “Multiple populations co-evolutionary particle swarm optimization for multi-

objective cardinality constrained portfolio optimization problem," Neurocomputing, vol. 430, pp. 58-70, Mar. 2021. doi:10.1016/j.neucom.2020.12.022.

[16] X. Zhang, Z. H. Zhan, and J. Zhang, "A fast efficient local search-based algorithm for multi-objective supply chain configuration problem," IEEE Access, vol. 8, pp. 62924-62931, Apr. 2020. https://doi.org/10.1109/access.2020.2983473.

[17] X. Liu, et al., "Coevolutionary particle swarm optimization with bottleneck objective learning strategy for many-objective optimization," IEEE Trans. Evol. Comput., vol. 23, no. 4, pp. 587-602, Aug. 2019. doi:10.1109/tevc.2018.2875430.

[18] H. Zhao et al., "Local binary pattern-based adaptive differential evolution for multimodal optimization problems," IEEE Trans. Cybern., vol. 50, no. 7, pp. 3343-3357 Jul. 2020. Doi:10.1109/tcyb.2019.2927780.

[19] Gunjan, S., et al. "Swarm intelligence approaches in supply chain management: potentials, challenges and future research directions. Supply Chain Management" An International Journal, 24(1), 107-123. (2019)./doi/10.1108/SCM-02-2018-0070.

[20] Hajikhani, Alborz, Mohammad Khalilzadeh, and Seyed Jafar Sadjadi. "A fuzzy multi-objective multi-product supplier selection and order allocation problem in supply chain under coverage and price considerations: An urban agricultural case study." Scientia Iranica 25.1: 431-449. 2018. DOI: 10.24200/sci.2017.4409.

[21] Loni, Parvaneh, Alireza Arshadi Khamseh, and Seyyed Hamid Reza Pasandideh. "A new multi-objective/product green supply chain considering quality level reprocessing cost." International Journal of Services and Operations Management 30.1: 1-22. 2018. doi:10.1504/ijsom.2018.091437.

[22] Cao, Cejun, et al. "A novel multi-objective programming model of relief distribution for sustainable disaster supply chain in large-scale natural disasters." Journal of Cleaner Production 174: 1422-1435. 2018. doi:10.1016/j.jclepro.2017.11.037.

[23] Singh, Sujeet Kumar, and Mark Goh. "Multi-objective mixed integer programming and an application in a pharmaceutical supply chain." International Journal of Production Research 57.4: 1214-1237. 2019. doi:10.1080/00207543.2018.1504172.

[24] Hendalianpour, Ayad, et al. "A linguistic multi-objective mixed integer programming model for multi-echelon supply chain network at bio-refinery." EuroMed Journal of Management 2.4: 329-355. 2018. doi:10.1504/emjm.2018.096453.

[25] E. Mastrocinque, B.Yuce , A. Lambiase and M. S. Packianather, "A Multi-Objective Optimization for Supply Chain Network Using the Bees Algorithm" International Journal of Engineering Business Management. Vol. 5, 38:2018. doi:10.5772/56754.

[26] R. E. Rasi, M.Sohanian "A multi-objective optimization model for sustainable supply chain network with using genetic algorithm" Journal of Modelling in Management. 3 August 2020. Doi: 10.30495/JSM.2022.1911221.1468.

[27] H. Banerjeea , Dr. V. Ganapathyb and Dr. V. M. Shenbagaramanc "Uncertainty Modelling in Risk-averse Supply Chain Systems Using Multi-objective Pareto Optimization" eprint arXiv:2004.13836 . April 2020.

[28] L. A Moncayo-Martínez, D. Zhang, "A Multi-objective Optimization for Supply Chain Network using Intelligent Water Drop" Industrial and Systems Engineering Research Conference. May (2014).

[29] S. Gupta1, A. Haq , I. Ali , B. Sarkar "Signifcance of multi objective optimization in logistics problem for multi product supply chain network nder the intuitionistic fuzzy environment" Complex & Intelligent Systems. 6 March 2021. https://doi.org/10.1007/s40747-021-00326-9.

[30] R.Sun, X. Wang, G. Zhao, "An Ant Colony Optimization Approach to Multi-Objective Supply Chain Model" The Second International Conference on Secure System Integration and Reliability Improvement . DOI 10.1109/SSIRI. 2008. doi:10.1109/ssiri.2008.35.

[31] P.Nguyen, K.Phuc, N. Thao "Ant Colony Optimization for Multiple Pickup and Multiple Delivery Vehicle Routing Problem with Time Window and Heterogeneous Fleets" Logistics, 10 May 2021. doi:10.3390/logistics5020028.

[32] Y.Wenfang, H.Guisheng, X. Pengcheng, L.Jingjing "Supply Chain Joint Inventory Management and Cost Optimization Based on Ant Colony Algorithm and Fuzzy Model" Technical Gazette 26, 1729-1737, 6,2019. doi.org/10.17559/TV-20190805123158.

[33] X. Zhang, , Z.Zhan, W. Fang, P. Qian, J. Zhang, "Multi-Population Ant Colony System with Knowledge-based Local Searches for Multiobjective Supply Chain Configuration" IEEE Transactions on Evolutionary Computation, DOI: 10.1109/TEVC.2021.3097339.

[34] A.Awad, R. Salem, H. Abdelkader , M. Abdul Salam" A Novel Intelligent Approach for Dynamic Data Replication in Cloud Environment " IEEE Access, 09 March 2021. doi:10.1109/access.2021.3064917.

[35] R. SALEM , M. Abdul Salam, H. Abdelkader ,A.Awad, " An Artificial Bee Colony Algorithm For Data Replication Optimization In Cloud Environments " Ieee Access, 03 December ,2019. DOI:10.1109/ACCESS.2019.2957436.

[36] A.Awad, R. Salem, H. Abdelkader , M. Abdul Salam" A Swarm Intelligence-based Approach for Dynamic Data Replication in a Cloud Environment " International Journal of Intelligent Engineering and Systems • March 2021.DOI:10.22266/ijies2021.0430.24.

Cylinder Liner Defect Detection and Classification based on Deep Learning

Chengchong Gao[1]
School of Mechanical Engineering
Nanjing Institute of Technology
Nanjing, China

Fei Hao[2]
Kangni Industrial Technology Research Institute
Nanjing Institute of Technology
Nanjing, China

Jiatong Song[3]
Kangni Industrial Technology Research Institute
Nanjing Institute of Technology
Nanjing,
China

Ruwen Chen[4]
Kangni Industrial Technology Research Institute
Nanjing Institute of Technology
Nanjing, China

Fan Wang[5]
Kangni Industrial Technology Research Institute
Nanjing Institute of Technology
Nanjing, China

Benxue Liu[6]
School of Mechanical and Power Engineering
Zhengzhou University
Zhengzhou, China

***Abstract*—The machine vision-based defect detection for cylinder liner is a challenging task due to irregular shape, various and small defects on the cylinder liner surface. To improve the accuracy of defect detection by machine vision a deep learning-based defect detection method for cylinder liner was explored in this paper. First, a machine vision system was designed based on the analysis of the causes and types of defects to obtain the field images for establishing an original dataset. Then the dataset was augmented by a modified augmentation method which combines the region of interest automatic extraction method with the traditional augmentation methods. Except for introduction of the anchor configuration optimization method, an XML file-based method of highlighting defect area was proposed to address the problem of tiny defect detection. The optimal model was experimentally determined by considering the network model, the training strategy and the sample size. Finally, the detection system was developed and the network model was deployed. Experiments are carried out and the results of the proposed method compared with those of the traditional methods. The results show that the detection accuracies of sand, scratch and wear defects are 77.5%, 70% and 66.3% which are improved by at least 26.3% compared with the traditional methods. The proposal can be used for field defect detection of cylinder liner.**

***Keywords*—Cylinder liner; defect detection; deep learning; machine vision**

I. INTRODUCTION

Cylinder liner is one of the most important parts of engine. Its surface quality will directly affect the working performance and service life of an engine. The surface quality will inevitably deteriorate due to the comprehensive effects of friction, high temperature and corrosion. If there are cracks, sand holes, air holes and other manufacturing defects in the cylinder liner itself, the degradation process will be greatly accelerated. Therefore, defect detection is of great significance in the production process of cylinder liner.

Machine vision inspection technology has the advantages of non-contact, easy to realize automation, and easy to analyze and process the detection results by computer which has been widely used in metal surface defect detection. At present, machine vision based defect detection methods mainly include traditional methods and deep learning methods [1-6]. The traditional machine vision defect detection method constructs the feature descriptors for different defects through image segmentation, feature extraction and other image processing algorithms. Through the descriptors, the surface defects are located, identified, graded, counted, stored and inquired. However, due to the influence of image quality, the complexity of industrial scene, the difference of defect shape and size, the traditional methods still often fails.

In recent years, deep learning based methods have been applied to defect detection of metal surfaces. However, there is no relevant research on how to apply deep learning to detect the surface defects of cylinder liners, and systematically describe the design and implementation of the detection system so far. Therefore, this paper will explore the deep learning based method for the cylinder liner surface defect detection and give the design process and its implementation of the defect detection system.

The main structure of this paper is as follows: the related works in this field are introduced in the second section; the defects types are analyzed and the design method of machine vision system is discussed in the third section; the optimal deep learning model are experimentally determined for cylinder liner defect detection in the fourth section; the design of detection system and the field experiments are given in the fifth section; the last section will summarize this paper.

II. Related Work

A cylinder liner defect detection system based on X-ray and linear array camera was built by Han Yueping of North China University and its key technologies were studied, such as sequence image filtering, threshold segmentation and morphological processing and calculation of defect parameter chain code tracking method [7, 8]. Considering that the probability of defects is low and the defect area is relatively small compared with the whole image, so the cylinder liner X-ray image is highly sparse. The group also proposed the use of compressed sensing algorithm for defect detection [9].

In other metal surface defect detection, the traditional machine vision method is more widely used. A multi-scale defect recognition method was proposed by Yu Jiahui et al. which has good accuracy and detection speed [10]. Tian Hongzhi et al. designed a micro defect detection system for grinding surface by combining plane illumination mode with multi angle illumination mode [11]. Mentouri zoheir et al. employed an improved dual cross algorithm to online monitoring of steel surface quality [12]. Aiming at the problems of complex defect pattern and low contrast between defect and background in steel strip surface defect detection, Liu Kun proposed a total variation image decomposition algorithm based on self-reference template and improved index gradient similarity [13]. Cao Binfang et al. proposed a defect detection method based on spatial-frequency multi-scale block local binary pattern to solve the problem of complex geometry and texture distribution of the nickel foam surface defect images [14]. Sun qianlai uses singular value decomposition to identify and locate surface defects of strip steel without image segmentation [15]. Based on the research on pseudo defect elimination, patch texture description and adaptive threshold segmentation, Liu Kun et al. proposed a new unsupervised steel surface defect detection model based on Haar-Weibull-variance [16]. Jeon Yong Ju et al. proposed the dual-light switching lighting technology to solve the problems of uneven brightness and various defects on the steel surface [17].

The core of traditional methods is to design and use feature descriptors, which include local binary pattern (LBP), histogram of oriented gradient (HOG), gray-level co-occurrence matrix (GLCM) and other statistical features. Feature descriptors are sensitive to lighting, background and other environmental factors. So it is very important to collect high-quality images. Through the optimal design of the imaging system, the difficulty of algorithm development can be reduced and the robustness of algorithm can be improved, but the cost of detection system must increase. Moreover, due to the complexity of industrial sites and the difference of defect shape and size, the failure of traditional algorithms still often occurs.

In recent years, with the successful application of deep learning in many fields of machine vision, more and more researchers are committed to using deep learning for defect detection in industrial field, aiming at improving the accuracy, efficiency, stability and reliability of the detection system. RetinaNet with difference channel attention and adaptively spatial feature fusion was proposed for steel surface defect detection by Cheng and Yu [18]. Zhang Jiaqiao et al. employed a CP-YOLOv3-dense neural network in the steel strip surface defect detection [19]. Xiao Ling et al. [20] proposed a surface defect detection method based on image pyramid convolution neural network model. Wei Rubo et al. [21] proposed a method for steel defect detection based on the fast regional convolution neural network. A steel surface defect detection model based on deformable convolution enhanced backbone network and pyramid feature fusion was proposed by Hao Ruiyang et al. [22].

These above deep learning based methods mainly focus on deep neural network (DNN) model. However, the dataset with support samples is more important than the DNN model. It is often difficult to obtain enough support samples for surface defect detection in industrial field. Therefore, how to improve the accuracy of deep learning-based defect detection method has become a research hotspot under the condition of a small number of samples.

A segmentation-based deep-learning architecture for the detection and segmentation of surface anomalies was proposed and demonstrated by Tabernik Domen et al. which can be trained with a small number of samples. In their experiments, only approximately 2530 defective training samples instead of hundreds or thousands were employed [23]. To address the problem that the existing defect datasets are generally unavailable for on-site deployment due to the limitation of data scale and defect types, Lv Xiaoming et al established a dataset named GC10-DET using a linear array image acquisition system to collect images [24]. To meet the challenge of detection the micro defect from high resolution images, a novel machine vision method was proposed for automatically identifying micro defects by Lian Jian et al. [25] and the main contributions of the proposal can be summarized as follows: 1) a defect exaggeration approach based on regularization, 2) a defect sample production method based on a generative adversarial network (GAN) and a convolutional neural network (CNN), and 3) a data augmentation method based on GAN. A novel approach for data augmentation was proposed by Jain Saksham et al. using GANs to create synthetic images to address the problem of time-consuming and high cost of on-site image acquisition. According to the comparative experiment, the performance of CNN architecture is significantly improved with GANs-based augmentation data and the sensitivity and specificity of the synthetically augmented CNN are 5.59% and 1.12% higher than those of the classical enhanced CNN, respectively [26].

The deep learning method has been employed in metal surface defect detection, but it has not been applied in cylinder liner surface defect detection. Therefore, this paper will take the lead in exploring the deep learning-based defect detection method for cylinder liner, designing a cylinder liner surface defect detection system, proposing a defect detection process, and giving a design case of cylinder liner surface defect detection system.

III. Defect Type Analysis and Machine Vision System

A. Defects Types of Cylinder Liner Surface

As shown in Fig. 1, the common surface defects of the cylinder liners are the sand defect, the crack defect, the wear defect, the oil defect, the scratch defect and the collision defect.

When casting cylinder liner, gas and non-metallic impurities can not be discharged before solidification of liquid metal, resulting in the formation of sand defects on the surface of cylinder liner after machining. The size of the sand defect is small, its contour is usually elliptical and its edge is smooth. Sand defect is one of the main defects of cylinder liner which may appear in any part of the cylinder liner. The existence of sand defects may greatly reduce the impact and fatigue resistances of cylinder liner which is easy to cause cylinder collapse, water leakage and other faults.

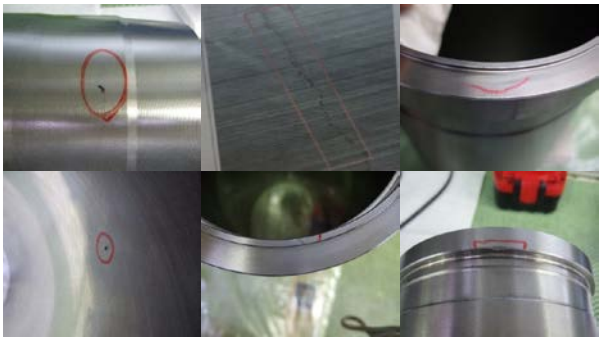


Fig. 1. Common Surface Defects of Cylinder Liner. From the Upper Left to the Lower Right, they are the Sand Defect, the Crack Defect, the Wear Defect, the oil Defect, the Scratch Defect and the Collision Defect.

In the process of machining, the cylinder liners were deformed under the combined action of various stresses. When the deformation exceeds the plastic limit, the slender flocculent or snowflake like cracks may appear on the surface of the cylinder liner. Most of the cracks occur on the inner and outer surface of the cylinder liner which may affect the reliability and replacement cycle of the cylinder liner.

Wear defect usually refers to drag, block or furrow deformation on the surface of cylinder liner during production and transportation. Compared with the normal area, the wear area is generally silvery white. Wear defects usually occur on the end face or outer surface of the cylinder liner which may reduce the sealing, corrosion resistance or wear resistance of the cylinder liner, resulting in the decrease of engine power.

Oil defect is a kind of pseudo defect which is formed by the air evaporation of the cleaning oil or antirust oil left on the outer or inner wall of the cylinder liner. Oil defects are charred black and long drop-shaped in appearance which are prone to false inspection.

Scratch defect is a kind of non-uniform and strip-shaped ravine defect which is caused by the friction between impurities and cylinder liner in the process of processing or transmission. Scratch defects often appear on the inner or outer wall of the cylinder liner which may lead to unreasonable fit clearance improper assembly and other problems, thus reducing the wear resistance and mechanical properties of the cylinder liner.

Collision defect are the falling off or blocky defects caused by the collision between the cylinder liner and the cylinder liner or the fence in the process of transportation. Most of them appear on the upper end face and skirt. Collision defects may reduce the cylinder liner sealing and engine efficiency, shorten the replacement cycle of cylinder liner, and lead to engine damage accidents.

In the field of cylinder liner surface quality inspection, there is no clear standard to identify the types and severity of the above six kinds of defects. However, the sand defect, scratch defect and wear defect are common in most enterprises. Therefore, the deep learning defect detection method with a small number of samples was studied to detect the above three types of defects.

B. Design of Machine Vision System

As shown in Fig. 2, the machine vision system for the cylinder liner defect detection consists of three area array cameras and a linear array camera. Camera 1, camera 2 and camera 3 are area array cameras. Camera 1 was employed to capture the image of top face and its object distance is about 270 mm. Camera 2 was employed to capture the image of the inner wall and its object distance is about 255 mm. The angle between its optical axis and the axis of camera 1 is about 62.5° ± 5°. Camera 3 was employed to capture the image of the skirt and its object distance is about 247 mm. The angle between its optical axis and the cylinder liner axis ranges from 26.5° to 36.5°. Camera 4 which is a linear array camera was used to capture the image of the outer wall and it object distance is between 338 mm to 358 mm.

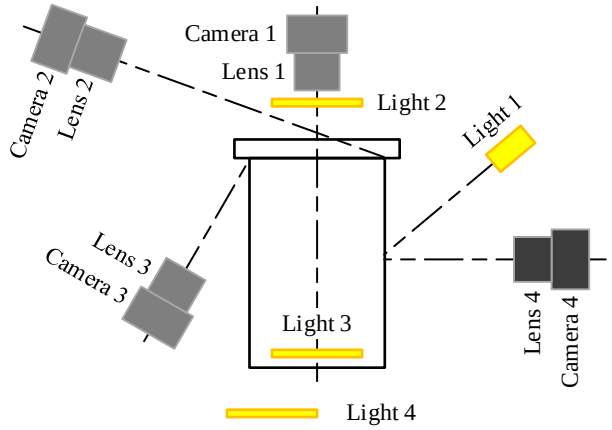


Fig. 2. Schematic Diagram of Machine Vision System for the Cylinder Liner Defect Detection.

The angle between the light 1 and the optical axis of the camera 4 is about 45°. The light 2 is a ring light source, the light 3 is a circular backlight and the light 4 is a ring light source of which the inner diameter is slightly larger than the outer diameter of the cylinder liner. Furthermore, the four light sources were installed at the determined positions.

With the above machine vision system, the images of the top face, skirt, inner wall and outer wall of the cylinder liner were collected, as shown in Fig. 3.

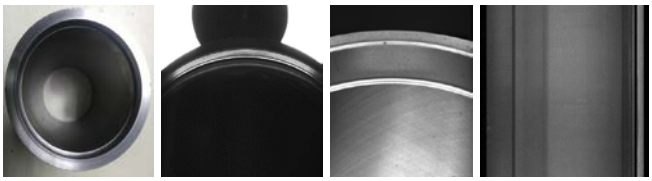


Fig. 3. Images of the Top Face, Skirt, Inner Wall and Outer Wall, Respectively.

IV. Optimal Depth Learning Model for Cylinder Liner Defect Detection

A. Establishment of Image Set

- Image Acquisition

There is no image set for cylinder liner surface defect detection. Therefore, 7500 images of cylinder liner were collected by using the above machine vision system with the guidance of field engineers and the image sizes are 2048 pixels × 2448 pixels. Among these images, 586 were defective. The sand defect images of the top face, the skirt, the inner wall and the outer wall of the cylinder liner are shown in Fig. 4, respectively.

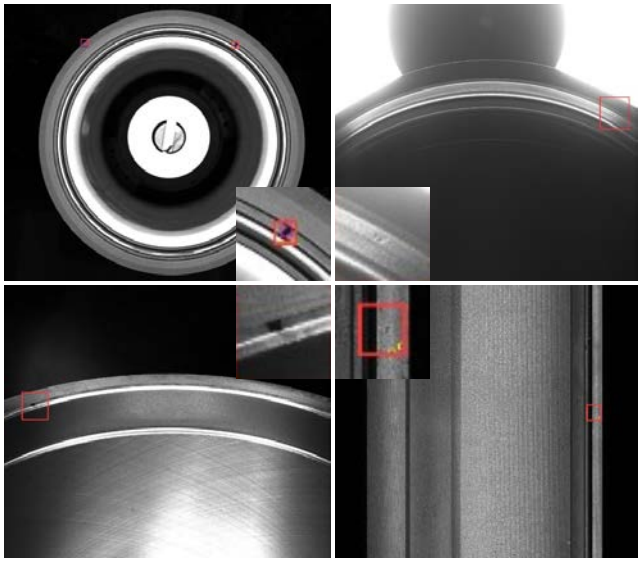


Fig. 4. Sand Defect Images of the Top Face, the Skirt, the Inner Wall and the Outer Wall of the Cylinder liner from Top Left to Bottom Right.

- Imageset Augmentation

The imbalanced datasets will make the results of convolutional neural network over biased to the classification of abnormal targets. In order to alleviate the over fitting problem and enhance the robustness of the network, considering that the cylinder liner image is gray image and the surface defects are small, we use upsampling method to expand the defect sample data to meet the requirements of neural network training data.

The common data augmentation methods mainly include rotation, offset, clipping and scaling. The rotation enhancements are accomplished by rotating the image to the right or left by 30°, 45°, 60° and 90°. The offset enhancements are to move the image around to change the original defect position. When the original image is converted in one direction, the remaining space can be filled with 0. The clipping enhancements are to cut the original image at 30°, 45°, 60° and 90° and fill the remaining space of the image with 0. The scaling enhancement makes the whole image scale in different ratios. In the convolution neural network training, more image invariant features can be learned to improve the detection accuracy.

The augmented dataset contains 5000 images which is basically balanced with the normal sample. The set was divided into the training set, the verification set and the test set by 8:1:1.

- Labelling Defects

Some regions of interest were extracted to reduce the search time of target region and the training time of neural network by the bi-dimensional maximum conditional entropy based threshold segmentation method of which a detailed derivation was carried out in our previous work. The mathematical model of grey entropy is as follows.

$$H(E \mid O) = -\sum_{i=0}^{s-1}\sum_{j=t}^{m-1} p_{ij} \log_2 p_{ij} \tag{1}$$

$$H(E \mid B) = -\sum_{i=0}^{n-1}\sum_{j=t}^{m-1} \tilde{p}_{ij} \log_2 \tilde{p}_{ij} \tag{2}$$

$$H(s,t) = \frac{1}{2}\big(H(E \mid O) + H(E \mid B)\big) \tag{3}$$

where m represents gradient levels of gradient image of gray image, n represents gray levels of gray image, p_{ij} represents the probability that a pixel with higher gradient but lower gray belongs to an edge and $\tilde{p}_{ij}$ represents the probability that a pixel with higher gradient and gray belongs to an edge.

(s*, t*) which makes the objective function H (s, t) take the maximum value is the optimal threshold for segmentation of a grayscale image and its gradient image.

t* was used to segment the cylinder liner image and extract the regions of interest such as the skirt and the top face. Labelimg which is a digital image labeling tool was employed to label the cylinder liner image. The label information is stored in an XML file which contains the image name, the image resolution, the defect size, the defect location and the defect names including the wear defect, the scratch defect and the sand defect and which is shown in Fig. 5.

B. Improvement of Anchor

A differential evolution search algorithm was employed to optimize the aspect ratios and scales of anchors to address the problem that the default anchor configuration turns out to be ineffective for detecting lesions of small size and large ratios [27].

Based on the default anchor configuration, this algorithm finds the optimal anchor setting of three scales and five ratios

through the iteration of the objective function. Suppose that the three scales are s_1, s_2, s_3 and $\varepsilon_1 > s_1, s_2, s_3 > \varepsilon_2 > 0$; the five ratios are β_2:1, β_1:1, 1:1, 1:β_1, 1:β_2 and $\varepsilon > \beta_2 > \beta_1 > 1$; where, ε_1, ε_2 and ε are constants. The optimal scales for detection of small size and large ratio objects are 0.680, 0.540 and 0.425 and the optimal scales are 3.27:1, 1.78:1, 1:1, 1:1.78 and 1:3.27. The anchor sizes remain unchanged which are still 32 pixels, 64 pixels, 128 pixels, 256 pixels and 512 pixels.

C. Determination of Optimal Deep Learning Model

Usually, the deep learning models were evaluated by the indicators of recall, precision and accuracy.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

```
<annotation verified="no">
  <folder>Desktop</folder>
  <filename>Image file name with suffix</filename>
  <path>Path name (including image file name)</path>
  <source>
    <database>Unknown</database>
  </source>
  <size>
    <width>An integer represents the defect width</width>
    <height>An integer represents the defect height</height>
    <depth>1</depth>
  </size>
  <segmented>0</segmented>
  <object>
    <name>Defect name</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>x-coordinate of defect box (minimum)</xmin>
      <ymin>y-coordinate of defect box (minimum)</ymin>
      <xmax>x-coordinate of defect box (maximum)</xmax>
      <ymax>y-coordinate of defect box (maximum)</ymax>
    </bndbox>
</annotation>
```

Fig. 5. XML File with Defect Information.

AP value was employed to represent the recognition rate of the deep learning model on a certain type of defect, which is equal to the area of the trapezoid enclosed by the *P-R* curve formed by recall and precision and the coordinate axes. mAP represents the average recognition rate of the model on all types of defects and its value ranges from 0 to 1. The larger the value of AP or mAP, the higher the defect recognition rate. The mAP will be adopted to evaluate and determine the optimal target detection model.

Three groups of experiments were carried out to analyse the influence of factors of the detection models, the number of datasets and the training strategy on convolutional neural network. The random gradient descent method was used to optimize the parameters, the loss function was calculated by the softmax layer, the initial learning rate is 0.00001 and the decay rate of the learning rate is 0.1. These configurations remained unchanged in the experiments to ensure comparability.

As shown in Fig. 6, the mAPs of SSD, Faster-RCNN and Retinanet with transfer learning strategy are higher than those of detection networks without transfer learning strategy. Therefore, the transfer learning strategy can shorten the training time and contribute to achieving higher detection accuracy which is suitable for a small number of samples.

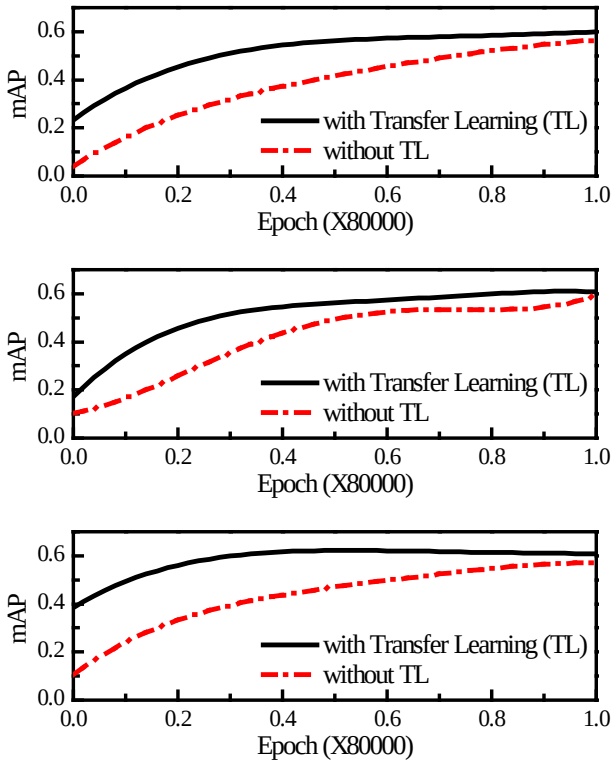


Fig. 6. Influence of Learning Strategy. (a), (b) and (c) is SSD, Faster-RCNN and Retinanet, Respectively.

The performance of the three networks with transfer learning strategy were compared and the comparison results are tabulated in Table I. According to the table, the mAPs of SSD, Faster-RCNN and RetinaNet are 0.597, 0.604 and 0.620, respectively. RetinaNet has the highest detection accuracy. In terms of time, the SSD is the fastest which takes 0.976 s and the RetinaNet is 0.212 s slower than the SSD. The RetinaNet will be employed as the main network for cylinder liner surface defect detection to achieve a good compromise in accuracy and speed.

TABLE I. PERFORMANCE COMPARISON OF THREE NETWORKS WITH TRANSFER LEARNING STRATEGY.

Detection Networks	Time (/s)	mAP
SSD	0.976	0.597
Faster-RCNN	1.360	0.604
RetinaNet	1.168	0.620

Furthermore, the effect of sample size on the performance of Retinanet was compared experimentally. Using 25%, 50%, 75% and 100% samples to train Retinanet, the mAP is 0.11, 0.40, 0.53 and 0.62, respectively. It can be seen that the more samples, the higher the accuracy of the network. Therefore, the data-driven deep learning model needs a lot of data to train its deep network, so as to obtain accurate feature extraction.

According to the above experiments, the Retinanet with the transfer learning strategy will be used for the cylinder liner defect detection.

D. Highlight Defect Areas of Support Samples

RetinaNet with transfer learning strategy achieves the highest mAP, but only 62%. By analyzing the dataset, it was found that the defect area only accounts for 0.25% of an image. However, the Retinanet uses the feature pyramid network (FPN) to produce feature maps with rich semantic. Most of the candidate image windows are background (negative classes) and only a few areas contain defects (positive classes). A large number of background cover up the defects which makes it impossible to fully extract the feature information of small defects in the process of the deep neural network training and finally leads to the low accuracy.

To highlight the defect to improve the detection accuracy, according to the location and area of the defect from the XML file, 100 pixels were extended to the top, bottom, left and right of the defect to form an image window to surround the defect. Then the defect area was updated by the formed window. Compared with the original image, the proportion of defect area was increased significantly and the detection accuracy is expected to be improved, as shown in Fig. 7.

The proportion of the redefined defect area in an image is significantly increased. Therefore, the accuracy of defect detection is effectively improved. Through the above method, the accuracy of Retinanet was improved from 0.62 to 0.71, which was increased by 14.52%. For the sand defect, scratch defect and wear defect, the AP value is 0.78, 0.69 and 0.66 respectively.

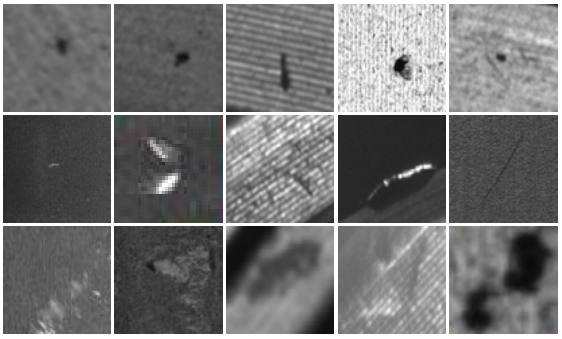


Fig. 7. Extended Defect Area. The First Line, the Second Line and the Third Line Respectively Correspond to Three Kinds of Defects: Sand, Scratch and Wear.

V. DESIGN OF DETECTION SYSTEM AND FIELD EXPERIMENT

A. Configuration of Detection System

The general layout of the cylinder liner surface defect detection system is shown in Fig. 8, the field oriented cylinder liner defect detection system is shown in Fig. 9, and the feeding system and image acquisition system are shown in Fig. 10.

According to Fig. 8 and Fig. 9, the workflow of the field oriented cylinder liner defect detection system is as follows: the cleaning subsystem cleans the cylinder liner; as shown in Fig. 10 (a), the feeding subsystem transports the cylinder liner to the end of the belt to trigger the position sensor; the lifting subsystem grabs and lifts the cylinder liner to the rotary platform accordingly; as shown in Fig. 10 (b), the rotary platform and the cylinder liner are sent to the detection room to complete the whole process of surface defect detection; they are sent out of the detection room and the sorting subsystem pushes the genuine and defective products to the corresponding conveyor belt according to the detection results.

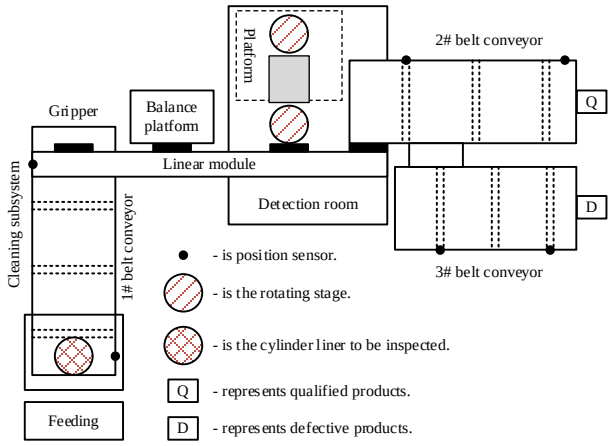


Fig. 8. Framework of the Detection System.

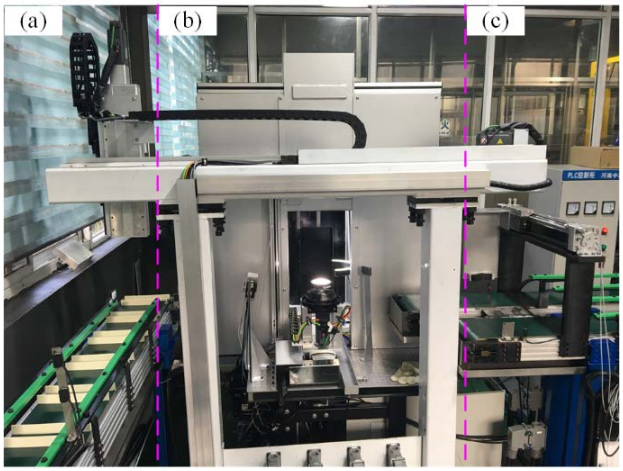


Fig. 9. On Site Cylinder Liner Defect Detection System. (a) is the Feeding Subsystem, (b) is the Detection Room and its Core Part is the Imaging Subsystem and (c) is the sorting Subsystem.

A workstation was used for network model training of which the memory is 256 GB, the model of the graphics card is NVIDIA Tesla P100, the video memory is 16 GB, and the operating system is Ubuntu 16.04. The framework of deep learning is tensorflow. The network model was offline trained by using the above dataset and it was deployed into an industrial-grade server successively.

B. Modular Design of Inspection System

After oil washed, the cylinder liner is easy to adhere to impurities such as lint and dust on the inner and outer walls which may easily lead to misjudgment during automatic optical inspection (AOI). Therefore, a cleaning subsystem was designed which mainly includes the air cylinder, air knife, oil buffer and other parts, as shown in Fig. 11 (a). The air knife forms an angle of 30 degrees with the axis of the cylinder liner to spray high-pressure gas to the surface of the cylinder liner to clean the surface of the cylinder liner and the air knife is pushed by the air cylinder to move up and down with the guidance of the guide rod to achieve the entire surface. The above cleaning process usually needs to be repeated twice. In addition, a shock absorber was employed at the joint between the cleaning device and the cylinder liner to reduce the impact of the start and stop impact of the air pump on the system.

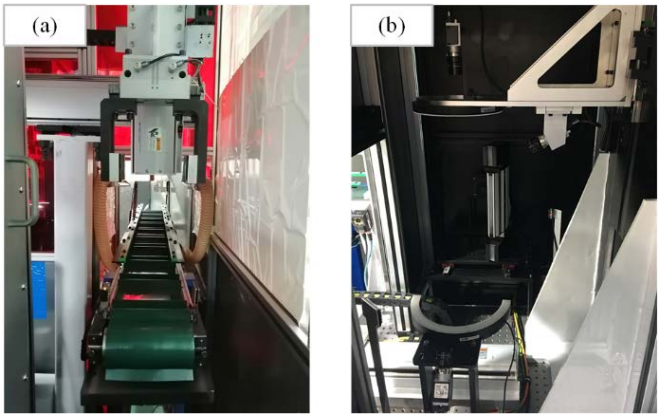


Fig. 10. Two subsystems. (a) is the Feeding Subsystem and (b) is the Imaging Subsystem.

The grasping device is the end executive device of the screw module of which the main design requirements include: (1) the clamping force should be large enough to ensure reliable clamping and avoid displacement or vibration during handling. However, it should not be too large to prevent the cylinder liner surface from being damaged. (2) The central line of the gripper coincides with the central line of the cylinder liner to ensure that the cylinder liner will not collapse during clamping to avoid secondary damage to the cylinder liner. (3) The gripper should be suitable for both D123 and D130 cylinder liners. Therefore, a gripper driven by air cylinder was designed. Oil resistant rubber was pasted on the inside of the claw to increase friction to prevent the cylinder from sliding, as shown in Fig. 11 (b).

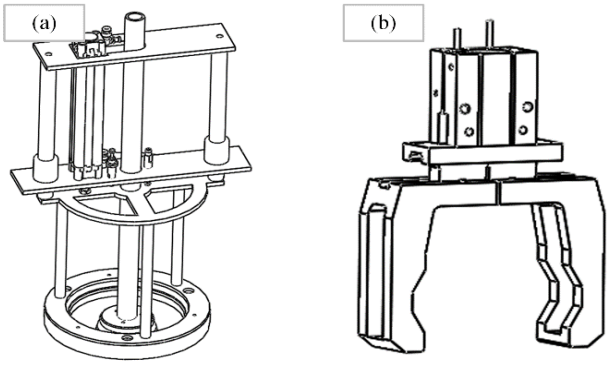


Fig. 11. Two submodular. (a) is the Cleaning Subsystem and (b) is the Grasping Device.

The design of detection room includes the design of rotating platform, as shown in Fig. 12 (a), and internal structure of the detection room. The cylinder liner is clamped on the rotating platform through the central positioning mode and the platform and the cylinder liner is driven to rotate to collect images of the inner and outer cylindrical surface to achieve the defect detection of the inner and outer walls. The higher the center positioning accuracy is, the greater the clamping force will be. If the clamping force is too large, it is easy to scratch texture on the inner surface, causing damage to the inner wall. If the positioning accuracy is too low, the cylinder liner cannot rotate reliably with the platform which affects the image acquisition. Therefore, the diameter of the rotating platform is 4mm smaller than the inner diameter of the cylinder liner and the motor drives the platform to rotate by the PLC. The internal structure design mainly includes the support and mechanical interface design for cameras, lights and rotating platform, as shown in Fig. 12 (b).

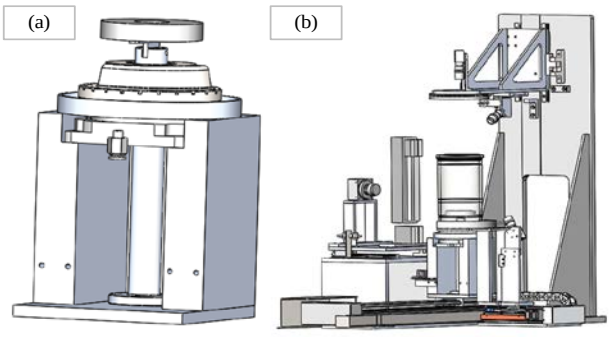


Fig. 12. Two Designs of Detection Room. (a) Rotating Platform and (b) Supports and Mechanical Interfaces.

VI. Experiments and Results

The field experiments were carried out using the network model trained in the third section and the dataset containing three types of defects: the sand, the scratch and the wear defect with 80 images for each type of defect. The experimental results were tabulated in Table II. The detection accuracy of sand defects is 77.5%, that of scratch defects is 70%, that of wear defects is 66.3% and the average accuracy is 71.26%.

TABLE II. Detection Results of Three Types of Defects

Defects	Images	Right	False detection	Undetected	Accuracy (%)
sand	80	62	13	5	77.5
scratch	80	56	15	9	70.0
wear	80	53	17	10	66.3

The comparison results of some traditional non-dedicated defect detection methods and the proposal were tabulated in Table III. The average accuracy of feature point registration-based method is 36.0% and that of morphology-based method is only 27.3%, but that of the method based on deep learning proposed in this paper is 71.3%. Compared with the feature point registration-based method, the proposed method improves the detection accuracies of sand, scratch and wear defect by 51.5%, 28% and 26.3%, respectively. Compared with the morphology-based method, the proposed method improves the detection accuracies of sand, scratch and wear defect by 51.5%, 44% and 36.3%, respectively. The deep learning-based method is more effective for cylinder liner defect detection compared with some traditional non-dedicated methods.

TABLE III. Comparison of the Proposal and Some Traditional Non-Dedicated Methods.

defects	Feature point based method	Morphology based method	Proposal
sand	26.0%	26.0%	77.5%
scratch	42.0%	26.0%	70.0%
wear	40.0%	30.0%	66.3%
mean	36.0%	27.3%	71.3%

The effect of proposed deep learning-based method was shown in Fig. 13. Our proposal can detect tiny defects such as the sand, the scratch and the wear defects, identify the types of defects and locate the defects in the very large cylinder liner images. The detection method can basically meet the actual cylinder liner surface detection requirements of the enterprise to continuously improve the product quality.

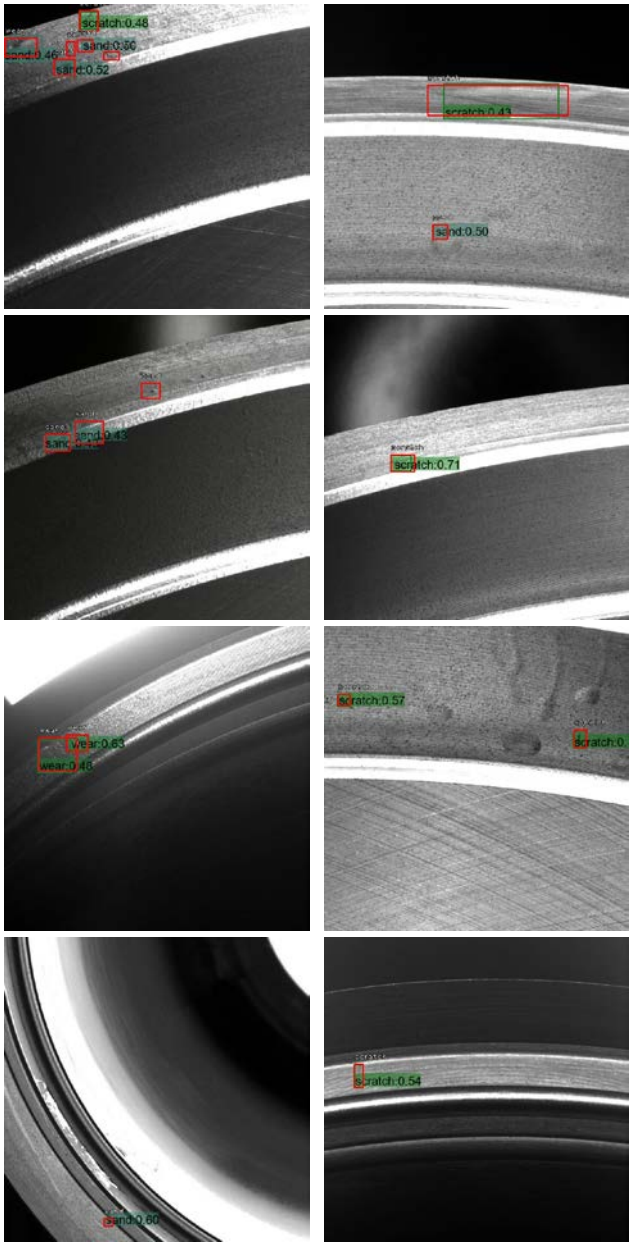


Fig. 13. Some Results of Proposed Deep Learning-Based Method.

However, the detection accuracy still needs to be improved. We can continue our research work from several aspects, such as the updated data sets [28], the motion platform control algorithms [29], the automatic labeling method [30], and the deep network model [31], which is expected to reduce the false detection rate and missing detection rate and further improve the accuracy of our method.

1) The network model can be further trained to improve its fitting accuracy by continuously adding the field data to the dataset to increase the amount of sample data.

2) The accuracy and reliability of the motion platform and its controller need to be further improved to ensure the acquisition of highly reliable images.

3) For the online defect detection, an automatic labeling method needs to be developed to avoid the missing labeling problem of manual labeling and improve the efficiency of defect labeling.

4) The network model can be improved and optimized to further improve its learning ability with a few samples and its detection ability for small defects.

VII. Conclusion

To address the actual needs of the cooperative enterprise, this paper developed a method and its system for cylinder liner surface defect detection based on deep learning. First, a machine vision defect detection system based on the causes and types of cylinder liner defects was built. Then, a dataset augmentation method based on the automatic extraction of region of interest was proposed which effectively increases the number of samples. Next, an automatic extension method of defect region was developed with the XML file which improves the detection ability of our proposal for small defects. After this, the network model and training strategy were experimentally determined and the influence of sample size on detection accuracy was discussed. Lastly, the scheme of implementing cylinder liner defect detection system in industrial field was given and the experiments were carried out. The results show that the detection accuracies of sand, scratch and wear defects are 77.5%, 70% and 66.3% which are improved by at least 26.3% compared with the traditional methods and that our method has achieved preliminary results and effects.

Acknowledgment

All authors express their gratitude to the Editor and the anonymous Reviewers for their valuable and constructive comments. And this research was supported in part by the National Natural Science Foundation of China (Grant No. 51705238) and the Promotion project for Modern agricultural machinery equipment and technology demonstration of Jiangsu Province (Grant No. NJ2021-58).

References

[1] Bhatt, P. M. et al., "Image-Based Surface Defect Detection Using Deep Learning: A Review," Journal of Computing and Information Science in Engineering, vol. 21, no. 4, pp. 1-23, 2021, doi: 10.1115/1.4049535.

[2] Tao, X., Hou, W., and Xu, D., "A Survey of Surface Defect Detection Methods Based on Deep Learning," Zidonghua Xuebao/Acta Automatica Sinica, vol. 46, no. x, pp. 1-18, 2020, doi: 10.16383/j.aas.c190811.

[3] Zhang, T., Liu, Y., Yang, Y., Wang, X., and Jin, Y., "Review of Surface Defect Detection Based on Machine Vision," Science Technology and Engineering, vol. 20, no. 35, pp. 14366-14376, 2020.

[4] Li, S., Yang, J., Wang, Z., Zhu, S., and Yang, G., "Review of Development and Application of Defect Detection Technology," Zidonghua Xuebao/Acta Automatica Sinica, vol. 46, no. 11, pp. 2319-2336, 2020, doi: 10.16383/j.aas.c180538.

[5] Qi, S., Yang, J., and Zhong, Z., "A Review on Industrial Surface Defect Detection Based on Deep Learning Technology," in 3rd International Conference on Machine Learning and Machine Intelligence, Virtual, Online, China, 18-20 Sept. 2020: Association for Computing Machinery, 2020, pp. 24-30, doi: 10.1145/3426826.3426832.

[6] Lu, R., Wu, A., Zhang, T., and Wang, Y., "Review on Automated Optical (Visual) Inspection and Its Applications in Defect Detection," Guangxue Xuebao/Acta Optica Sinica, vol. 38, no. 8, pp. 1-36, 2018, doi: 10.3788/AOS201838.0815002.

[7] Guo, J., "Research on Defect Detection Technology of the Cylinder Liner Based on Line Array Imaging," Master, North University Of China, Taiyuan, China, 2017.

[8] Li, H., "Research on the Technology of Automatic Classification of X-ray Cylinder Liner Defect Based on Grey Theory," Master, North University Of China, Taiyuan, China, 2017.

[9] Zhuo, H., Han, Y., and Guo, J., "Study on Automatic Defect Detection Technology of the Cylinder Liner Based on X-ray," Techniques of Automation and Applications, vol. 37, no. 02, pp. 93-96+106, 2018.

[10] Yu, J., Gao, H., Sun, J., Yang, W., Jiang, Y., and Ju, Z., "Automatic and efficient metallic surface defect detection based on key pixel point locations," IEEE Sensors Journal, 2020, doi: 10.1109/JSEN.2020.3017737.

[11] Tian, H., Wang, D., Lin, J., Chen, Q., and Liu, Z., "Surface defects detection of stamping and grinding flat parts based on machine vision," Sensors (Switzerland), vol. 20, no. 16, pp. 1-17, 2020, doi: 10.3390/s20164531.

[12] Mentouri, Z., Doghmane, H., Moussaoui, A., and Bourouba, H., "Improved cross pattern approach for steel surface defect recognition," International Journal of Advanced Manufacturing Technology, vol. 110, no. 11-12, pp. 3091-3100, 2020, doi: 10.1007/s00170-020-06050-x.

[13] Liu, K., Luo, N., Li, A., Tian, Y., Sajid, H., and Chen, H., "A New Self-Reference Image Decomposition Algorithm for Strip Steel Surface Defect Detection," IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 7, pp. 4732-4741, 2020, doi: 10.1109/TIM.2019.2952706.

[14] Cao, B., Li, J., and Qiao, N., "Nickel foam surface defect detection based on spatial-frequency multi-scale MB-LBP," Soft Computing, vol. 24, no. 8, pp. 5949-5957, 2020, doi: 10.1007/s00500-019-04513-2.

[15] Sun, Q., Wang, Y., and Sun, Z., "Rapid surface defect detection based on singular value decomposition using steel strips as an example," AIP Advances, vol. 8, no. 5, pp. 1-12, 2018, doi: 10.1063/1.5017589.

[16] Liu, K., Wang, H., Chen, H., Qu, E., Tian, Y., and Sun, H., "Steel Surface Defect Detection Using a New Haar-Weibull-Variance Model in Unsupervised Manner," IEEE Transactions on Instrumentation and Measurement, vol. 66, no. 10, pp. 2585-2596, 2017, doi: 10.1109/TIM.2017.2712838.

[17] Jeon, Y., Choi, D., Lee, S., Yun, J., and Kim, S., "Steel-surface defect detection using a switching-lighting scheme," Applied Optics, vol. 55, no. 1, pp. 47-57, 2016, doi: 10.1364/AO.55.000047.

[18] Cheng, X. and Yu, J., "RetinaNet with Difference Channel Attention and Adaptively Spatial Feature Fusion for Steel Surface Defect Detection," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-11, 2021, doi: 10.1109/TIM.2020.3040485.

[19] Zhang, J., Kang, X., Ni, H., and Ren, F., "Surface defect detection of steel strips based on classification priority YOLOv3-dense network," Ironmaking and Steelmaking, 2020, doi: 10.1080/03019233.2020.1816806.

[20] Xiao, L., Wu, B., and Hu, Y., "Surface Defect Detection Using Image Pyramid," IEEE Sensors Journal, vol. 20, no. 13, pp. 7181-7188, 2020, doi: 10.1109/JSEN.2020.2977366.

[21] Wei, R., Song, Y., and Zhang, Y., "Enhanced faster region convolutional neural networks for steel surface defect detection," ISIJ

International, vol. 60, no. 3, pp. 539-545, 2020, doi: 10.2355/isijinternational.ISIJINT-2019-335.

[22] Hao, R., Lu, B., Cheng, Y., Li, X., and Huang, B., "A steel surface defect inspection approach towards smart industrial monitoring," Journal of Intelligent Manufacturing, 2020, doi: 10.1007/s10845-020-01670-2.

[23] Tabernik, D., ela, S., Skvar, J., and Skoaj, D., "Segmentation-based deep-learning approach for surface-defect detection," Journal of Intelligent Manufacturing, vol. 31, no. 3, pp. 759-776, 2020, doi: 10.1007/s10845-019-01476-x.

[24] Lv, X., Duan, F., Jiang, J., Fu, X., and Gan, L., "Deep metallic surface defect detection: The new benchmark and detection network," Sensors (Switzerland), vol. 20, no. 6, pp. 1-15, 2020, doi: 10.3390/s20061562.

[25] Lian, J. et al., "Deep-Learning-Based Small Surface Defect Detection via an Exaggerated Local Variation-Based Generative Adversarial Network," IEEE Transactions on Industrial Informatics, vol. 16, no. 2, pp. 1343-1351, 2020, doi: 10.1109/TII.2019.2945403.

[26] Jain, S., Seth, G., Paruthi, A., Soni, U., and Kumar, G., "Synthetic data augmentation for surface defect detection and classification using deep learning," Journal of Intelligent Manufacturing, 2020, doi: 10.1007/s10845-020-01710-x.

[27] Zlocha, M., Dou, Q., and Glocker, B., "Improving RetinaNet for CT Lesion Detection with Dense Masks from Weak RECIST Labels," in 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13-17 Oct. 2019, vol. 11769: Springer Science and Business Media Deutschland GmbH, pp. 402-410, doi: 10.1007/978-3-030-32226-7_45.

[28] Tang, C., Zhu, Q., Wu, W., Huang, W., Hong, C., and Niu, X., "PLANET: Improved Convolutional Neural Networks with Image Enhancement for Image Classification," Mathematical Problems in Engineering, vol. 2020, p. 1245924, 2020/03/11 2020, doi: 10.1155/2020/1245924.

[29] Meng, C., Shi, J., Hao, F., and Li, P., "Error Calibration Method Based on Perspective Mapping for Wafer Automatic Optical Inspection System," IEEE Transactions on Instrumentation and Measurement, vol. 71, 2022, doi: 10.1109/TIM.2022.3150567.

[30] Cheng, Q., Zhang, Q., Fu, P., Tu, C., and Li, S., "A survey and analysis on automatic image annotation," Pattern Recognition, vol. 79, pp. 242-259, 2018/07/01/ 2018, doi: 10.1016/j.patcog.2018.02.017.

[31] Wang, S., Wu, T., Zheng, P., and Kwok, N., "Optimized CNN model for identifying similar 3D wear particles in few samples," Wear, vol. 460-461, p. 203477, 2020/11/15/ 2020, doi: 10.1016/j.wear.2020.203477.

Mobile Payment Transaction Model with Robust Security in the NFC-HCE Ecosystem with Secure Elements on Smartphones

Lucia Nugraheni Harnaningrum[1]
Department of Information Technology, Indonesia Digital Technology University, Yogyakarta, Indonesia

Ahmad Ashari[2], Agfianto Eko Putra[3]
Department of Computer Science and Electronics
Gadjah Mada University, Yogyakarta, Indonesia

***Abstract*—The Near Field Communication embedded (NFC-embedded) smartphone consists of two ecosystems, namely Near Field Communication Subscriber Identity Module Secure Element (NFC-SIM-SE) and Near Field Communication Host Card Emulation (NFC-HCE). NFC-SIM-SE places secure elements in smartphones, while NFC-HCE places secure elements in the cloud. In terms of security, the location of secure elements in the cloud is one of the weaknesses of NFC-HCE. The APL-SE transaction model is developed as a solution to improve transaction security with NFC-enabled mobile. This model moves the secure elements of the NFC-HCE ecosystem from the cloud to the smartphone so that when the transaction is made, the smartphone does not communicate with the outside network to access the secure element. The APL-SE transaction model is tested using dummy data to calculate the processing time measurements for each step. The model is also tested for the encryption process. The encrypted data is compared with the original data, then the randomness is calculated. This transaction model is also tested by looking at the data randomness, which shows that the encrypted data is declared random. Random data increases data security. The transaction model test shows that the transaction runs well because the encrypted data is proven random, and the execution time is 1,074 ms. The time of 1,074 ms is far below an attacker's time to decipher the encrypted data. Random and fast encryption results indicate that transactions are secure. This achievement makes the opportunity for attackers to manipulate data small, so security is increased.**

***Keywords*—Transaction; near field communication; mobile; secure element; encryption**

I. Introduction

Smartphones have become tools and facilities that are used daily by people in general. The use of smartphones is not only for communication itself. It is also used for other purposes such as accessing news, doing office work with applications that can be installed on smartphones, and even for business transactions. According to [1], the development of non-cash transactions is increasing from time to time. Nowadays, mobile transaction uses digital wallets. It is already widely used by banks and non-banks [2].

Payment transactions using Near Field Communication (NFC) are an alternative digital payment method. Payments with NFC have advantages in transaction speed, secure storage of card data networks can be deleted remotely [3]. The implementation of devices for transactions with NFC has been carried out, including the existence of a payment authorization system using Near Field Communication - Radio Frequency Identification (NFC-RFID) devices [4]. Payment systems using a cam-wallet also use NFC to communicate with merchants [5]. Smartphones with NFC hardware were around 64 percent in 2018, and NFC-enabled Point of Sales (POS) reached 53 percent globally in 2007 [6].

NFC-enabled mobile communication has two ecosystems [7]. These ecosystems are NFC Subscriber Identity Module Secure Element (SIM-SE) and Host Card Emulation (HCE). NFC SIM-SE performs transactions without an internet connection because Secure Element (SE) is on the smartphone. NFC-HCE requires an internet connection because SE is in the cloud. This condition makes security a significant problem in using NFC-HCE, so appropriate actions are needed to protect payment security. SE being moved to the cloud causes the need to send credential keys from the cloud to the device. However, the NFC-HCE ecosystem that does not need to use a SIM in its implementation can be an advantage and will be more widely applied in the future. NFC-HCE is also a solution in several countries' NFC mobile payment systems, which enforces SIMs produced without SE.

Therefore, this study develops an application-based NFC-HCE Model to optimize the performance of NFC-HCE in a mobile payment system to allow the use of a SIM with or without an SE to securely make NFC mobile payment transactions. The model implements SE in the NFC-SIM-SE ecosystem into the NFC-HCE ecosystem and stays in the smartphone. This model allows communication between devices (smartphones and NFC readers) by implementing a transaction system with the support of a security system. Overall, the model consists of two main parts: initialization and transactions. The initialization stage involves preparing an NFC-enabled smartphone for transactions, namely by registering to the server, both users (smartphones) and cards (one user can have many cards). The initialization model ensures that the credential data is stored in the smartphone safely and verified [8]. The transaction stage contains transactions between NFC from the card owner's smartphone to the point of sales (POS). In this study, the discussion focuses on developing the transaction section, which sends secure data between the cardholder's smartphone and POS, namely APL-SE.

The result of this research is a model that can be developed into a payment system protocol. The model can be used on a small or large scale of payment. Sellers and buyers must have a smartphone that has NFC-HCE facilities. Payment systems at small and large shops that use online payment systems can use APL-SE as a means.

The structure of this paper is as follows. After this introduction, it will present the research on mobile payment systems that have been carried out before. Then it is connected with the study that will be made and its novelty analyzed. The following section discusses the proposed mobile payment system model. This proposed model performs transactions in the NFC-HCE ecosystem that does not connect to the internet because the secure elements are already stored in the smartphone. Thus, the model reduces the risk of data being misused due to the smartphone connected to the server via a public internet network. This model can prevent attacks that occur because of data retrieval at the time of use and stored for use at another time. The model can also prevent attacks that may occur because the retrieved data is used at other transaction times. Then, the model is tested and analyzed to get a conclusion whether the model can prevent the attack.

II. Related Work

Badra [9] uses a user identity accompanied by a random value for authentication. The Badra model uses the NFC-SE ecosystem, which emulates the card. The solution to prevent NFC-related attacks is using certificate-based authentication between PoS and TTP and shared-secret-based authentication between TTP and NFC-enabled devices. The assumption is that the secret key shared between the TTP and the mobile is securely stored in the SE. Its cryptographic calculations are also performed inside the SE. The Badra model uses five stages, is simpler, and can overcome some of the possible attacks.

Poughomi and Grønli et al. [7] proposed the NFC protocol to be implemented on the Web of Things (WoT). This protocol is divided into two parts; the authentication and transaction sections. Transactions from a merchant terminal (POS) to a communication smartphone use NFC and from a smartphone to an MNO using a GSM line with SMS. Currently, the model of communication with SMS done several times is not so popular anymore.

Nashwan [10] proposed a secure authentication protocol for NFC (SAP-NFC). It is a protocol to overcome replay attacks, impersonate attacks, to track attacks, and desynchronize attacks using the registration and authentication phases. This security model uses a hash function and multiple authentications. The SAP-NFC protocol is more likely to overcome attacks because authentication is carried out for each data exchange.

Cossman and Liu [11] proposed two authentication steps. This research shifts the protection of user data to a user-centered approach. The first authentication uses the system keystore, and the second uses a passcode. Meanwhile, the two-factor authentication proposed by Munch-Elinsen et al. [12] is by providing the user with a PIN code twice. This authentication is divided into five phases. The PIN is entered in the first phase, where the user enters the PIN into the application, and in the fifth phase, the user also enters the PIN again to be verified by the POS. The two proposals for two-factor authentication are attempts to secure transactions. Cossman's proposal is simpler because it does not require other parties, while Munch-Elinsen's proposal requires other parties to send and receive SMS.

Cryptography is used to design security tags by consuming less energy. The cryptographic methods used are Asymmetric Cryptography and Symmetric Cryptography; encryption and decryption using Advanced Encryption Standard (AES-128), and digital signature generation using elliptic curve digital signature algorithm. Özcanhan et al. [13] proposed The EKATE protocol. The protocol uses asymmetric encryption for secure data communication. The AES algorithm is used to encrypt the data to be sent to and from the tag.

The relay attack scenario is made on a system that involves smartcards; smartphones on both Relay attacks attack smartphones by changing peer-to-peer mode into Card Emulation mode, then reading the smartcard. The message is captured and forwarded to the proxy device, and the device sends the message as if it were the real owner of the smartphone. These relay attacks occur at the application layer, so they must take real-time countermeasures. Dang Zhou et al. [14] studied relay attacks on NFC by analyzing the weaknesses of ISO / IEC 14443-4 when facing relay attacks. This drawback appears to be quite common to all types of AFC systems that follow this standard globally. Then an experimental relay method was designed and carried out a relay attack. The results show that the protocol is vulnerable to attack. Two counterattacks are also proposed and discussed the feasibility and practicality of these countermeasures. The results show that the attacks carried out successfully generated delays during transactions. Sujithra et al. [15] proposed a data encryption protocol to be stored on a smartphone with three tiers tested in local smartphone and cloud environments. The test results show that in terms of the speedup ratio, the combined algorithm AES+MD5+ECC is better, and the AES algorithm is better in terms of average processing time.

Al-Fayoumi and Nashwan [16] use the registration stage to ensure that NFC devices are registered in the AuC database. This registration stage uses four steps: sending a registration request message containing identity and a random number, AuC generates a secret key based on the parameters in the request message, and a confirmation message is sent back to the NFC device by the AuC. The NFC device executes a derivative function to obtain the secret keys. The authentication phase is the second stage of the secure authentication protocol for NFC mobile payment systems (SAP-NFC). POS and mobile NFC carry out initial authentication. POS generates random numbers and sends them to NFC mobile. NFC Mobile uses this random number to initiate the authentication challenge message. After the message is processed using these parameters, it is sent to the NFC POS. AuC verifies NFC devices with a set of identity and authentication parameters. The design of the SAP-NFC protocol fulfills the following requirements. First, all parties involved in authentication can generate random numbers. Second, the AuC and the NFC device can update the secret key for each authentication session. Third, both the old secret key from the previous

authentication session and the new secret key from the current authentication session of the NFC device will be stored in the AuC database. Fourth, mutual authentication must be carried out between all parties involved in authentication. Fifth, the KDF function (Derivation function) derives the new session key. Sixth, the identity of the party performing the authentication is hidden by a series of hash functions. As a result, the SAP-NFC protocol claims to achieve the highest level of security with mutual authentication, forward/backward secrecy, anonymity, and untraceability. The SAP-NFC protocol can also defeat attacks such as replay attacks, impersonation attacks, tracking attacks, and desynchronization attacks.

Alatar and Achemlal [17] stated that pure HCE has not been trusted for payments but has attracted the attention of Visa and Master cards. It still needs efforts to earn that trust. Asaduzzman et al. [18] stated that NFC is suitable for IoT devices. This paper discusses the protocol for sending data in NDEF format. This protocol uses certificates. Transaction data security is carried out in several ways. The first way is a modified certificate. This method is initiated by requesting a certificate via a handshake and will terminate the transaction if the certificates are not the same. The second way is with modified data. If the signatures do not match, it means that there is a modification to the data, then the data is discarded. There is a hash code matching mechanism between the message and the signed hash. Jamming attacks can be detected in the presence of interference. So, if there is interference, data is prohibited from being stored.

Alzahrani [19] identified an attack with a reapplication tag with TRD by tracking how many times the tag was read. The tag should only be read once when the goods are distributed and reach their destination. The original tag is already considered a second reading if the verified tag is fake. Pourghomi, Piere, et al. [7] developed an NFC protocol that begins with user authentication. You do this by checking the validation by checking the PIN. Then the data is stored in the tamper-resistance chip. All of that is done in the NFC-HCE ecosystem. Wenxing [20] created an NFC communication mechanism to prevent eavesdropping using electronic circuits. As a result, it can reduce the threat of eavesdropping. Fan et al. [21] created a protocol beginning with initialization. At the initialization stage, a pseudorandom generator and key are generated. Then put both in the valid and legitimate reader tags. At the tag identification stage, the reader generates a random timestamp and random number and sends an authentication to the tag. The time received must be greater than the time to transfer, otherwise, authentication is not successful. From the security side, the attack is detected with a timestamp for an anti-DoS attack scheme.

Nour et al. [22] overcome Replay attacks and Man in the Middle Attacks that often occur in NFC-enabled mobile transactions by creating a security protocol. This security protocol is used for mobile transactions. Prevention of replay attacks is done by using random numbers and timestamps. Meanwhile, the prevention of the Man in the Middle Attack is prevented by mutual authentication between the client's payment device and the small merchant's NFC smartphone. This payment architecture can also facilitate mobility because it uses mobile devices and is secure because the proposed protocol can solve EMV vulnerabilities without changing EMV principles.

The use of cards for travel has been used, one of which is AFC payments using LessPay. Fan Dang et al. [23] research shows that tampering with entrance data and relay attacks on AFC cards must be watched. This study simulates the attack and provides a solution for its prevention.

The NFC transaction security protocol was proposed by Ali Al-Haj et al. [24]. This protocol is successful in preventing malicious network attacks such as the impersonation and replay attack, the session key security attack, the brute force attack, and the Man-in-the-middle attack. These attacks are prevented by mutual authentication, non-repudiation of transaction messages, data integrity, confidentiality, data privacy, and validity of Banking Data. This protocol involves actors. The first actor is NFC-enabled mobile. This device is the main device in this mobile payment. This NFC-enabled mobile will communicate with the Management Authentication Server (MAS) to obtain a session key. The second actor is POS. This POS communicates with NFC-enabled mobile for transactions and with the issuing bank (BI) via a secure channel (TLS). The third actor is BI. BI communicates with POS to verify Mobile payments in online EMV transaction mode. The fourth actor is MAS. MAS provides management and authentication for secure mobile payment transactions.

Prevention of attack by brute force was carried out by Madhoun, M et al. [25]. On payment processing, each transaction takes about 500ms; therefore, to achieve this attack, one needs to have access to the client's payment device for 38 minutes. Although NFC is claimed to be able to communicate at short distances (only up to 10 cm), it has been proven that relay attacks can carry out attacks when communication occurs with NFC. The attack is carried out by adding an amplifier to extend the reading distance. This attack can retrieve credential data and use it for online transactions at another time. Authentication to the client is carried out with a PIN that is verified with the PIN data stored on the server in two ways, namely online verification with symmetric encryption and offline with an asymmetric key to the issuing bank or by comparing the PIN stored in the memory of the client's payment device. Verification can also be done with a signature by the client or without the Cardholder Verification Method (CVM). Verification without CVMini is used for fast payments and in small nominal amounts.

The transaction begins by entering the username and password. If the authentication server identifies as a valid user, the server will send a one-time password (OTP). The transaction will time out after a particular time. The user will be notified if the attacker cannot complete the payment process within the specified timeframe (30 seconds for Apple Pay). It not only places a time limit on the attacker but also raises consent issues [26].

Security and user trust issues are issues that are widely discussed in research on NFC-based mobile payment systems. Protocols or models have been built trying to overcome these problems. However, existing models still have gaps in vulnerability to attacks. One of them is an attack carried out during the transaction because of the communication from the

smartphone to the financial institution server. This study reduces this vulnerability by creating a model that does not require communication between smartphones and financial institution servers during smartphone transactions with POS using NFC-enabled mobile. It also has advantages in the amount of data exchange and the absence of communication from smartphones to financial institutions during transactions. This situation occurs in the HFC-HCE ecosystem.

III. Proposed Method

The proposed NFC mobile payment system consists of two stages: initialization and transaction, and this paper focuses on developing the transaction stage. The payment card is declared safe to be stored on the smartphone during the initialization process. The payment card is ready to be used for transaction processing. The transaction model ensures that transactions between the cardholder's smartphone and the POS are safe and correct. The architecture of the proposed model is shown in Fig. 1.

This transaction model puts the security system in the HCE ecosystem into smartphones. A security system is created in the form of an application that will ensure data security and communication between smartphones and POS. The security system refers to the SE hardware inside the SIM card in the NFC-SIM-SE ecosystem. Element identification is made by synthesizing NFC-SIM-SE elements and NFC-HCE elements in the cloud into application elements. Elements in both ecosystems are analyzed and then adapted to security system applications.

This transaction model is a model for data usage. In this model, data security is carried out only when communication is carried out using NFC-HCE. Two things will be prevented. The first is to make data safe from attackers, and secondly, when there is a data request, the data will be sent to the right place. Secure element created we will refer to as APL-SE. APL-SE on smartphones was used for transactions, and at that time, three entities were involved, namely smartphones, POS, and Financial Institutions. This study focuses on the security of transactions between smartphones and POS.

The smartphone and the POS transaction model are carried out with NFC-HCE communication in card emulation mode. In this mode, when a transaction is made, the smartphone and the POS communicate by acting as the initiator and target. When sending data, the smartphone or POS acts as the initiator and the others as the target. This model is shown in Fig. 2.

The first step, POS as a target, sends data on the number of purchases to the smartphone. Currently, POS is the target, and the smartphone is the initiator. The smartphone activates the HCE service and starts the transaction by sending a connection request to the POS. The smartphone taps it for this delivery. The data sent is String data which contains information on the payment amount. This process is shown in Fig. 3.

The transaction model created to enable these two devices enables the HCE service. In the beginning, POS recorded transaction data, then saved the data to a shared preference variable. The smartphone sends a request as an APDU command to the POS, and the POS responds by providing a response that is processing the APDU command. This response brings the amount of transaction data prepared previously by the POS.

In the second step, the smartphone receives the data, then activates the SE software. Notification of approval to POS is done by sending card data to POS. The smartphone verifies the request from the POS, then activates the smartphone application. The smartphone application requests a pin from the user, and this pin is stored in the data stored on the smartphone during the initialization process. When a pin is entered, it is matched utilizing a pin stored in an encrypted state, decrypted first, and then matched with the input data. If the pin matches the data, the process is continued. Otherwise, it is closed. Currently, smartphones are the target, and POS is the initiator. If the pin is correct, an RSA encryption key (KeyRSA) will be generated. The selected card is retrieved data, then stored in a JSON Object (JSON (Dcard Key data and files, which are already stored during the initialization process, are encrypted (Enkr(Dtrans)). This encrypted transaction data is sent to the POS (ED).

In the third step, after the POS receives the card data, the POS verifies the card data to the financial institution server and gets an approval notification if the data matches. POS as the initiator and financial institutions as the target, the data sent from the smartphone is received by the POS and then decrypted. Previous data plus payment data.

$$Dtrans = Duser + Dpay \quad (1)$$

The data is encrypted by the POS and sent to the Financial Institution.

$$Etrans = E(RV, Dtrans) \quad (2)$$

In Financial Institutions, data is decrypted and matched with customer data. The authorization code is encrypted (OT) if the Financial Institution is verified. If it is not verified, an unverified message is sent. Financial Institutions send data or notifications to POS.

In the fourth step, the POS executes the payment and sends a notification to the smartphone if the transaction has been successful. POS is the initiator, while the smartphone is the target. Authentication data and smartphone data are decrypted, as are user data decrypted. Next, the payment process is carried out. The process is completed by sending a notification to the smartphone. Currently, POS is the initiator, and the smartphone is the target.

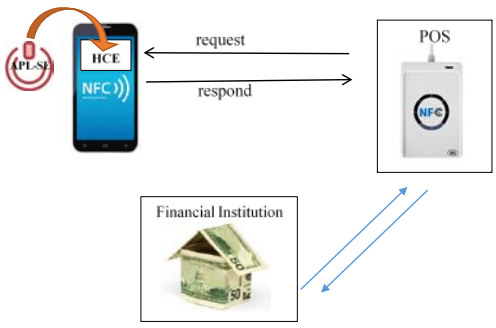


Fig. 1. APL-SE Architectural Design.

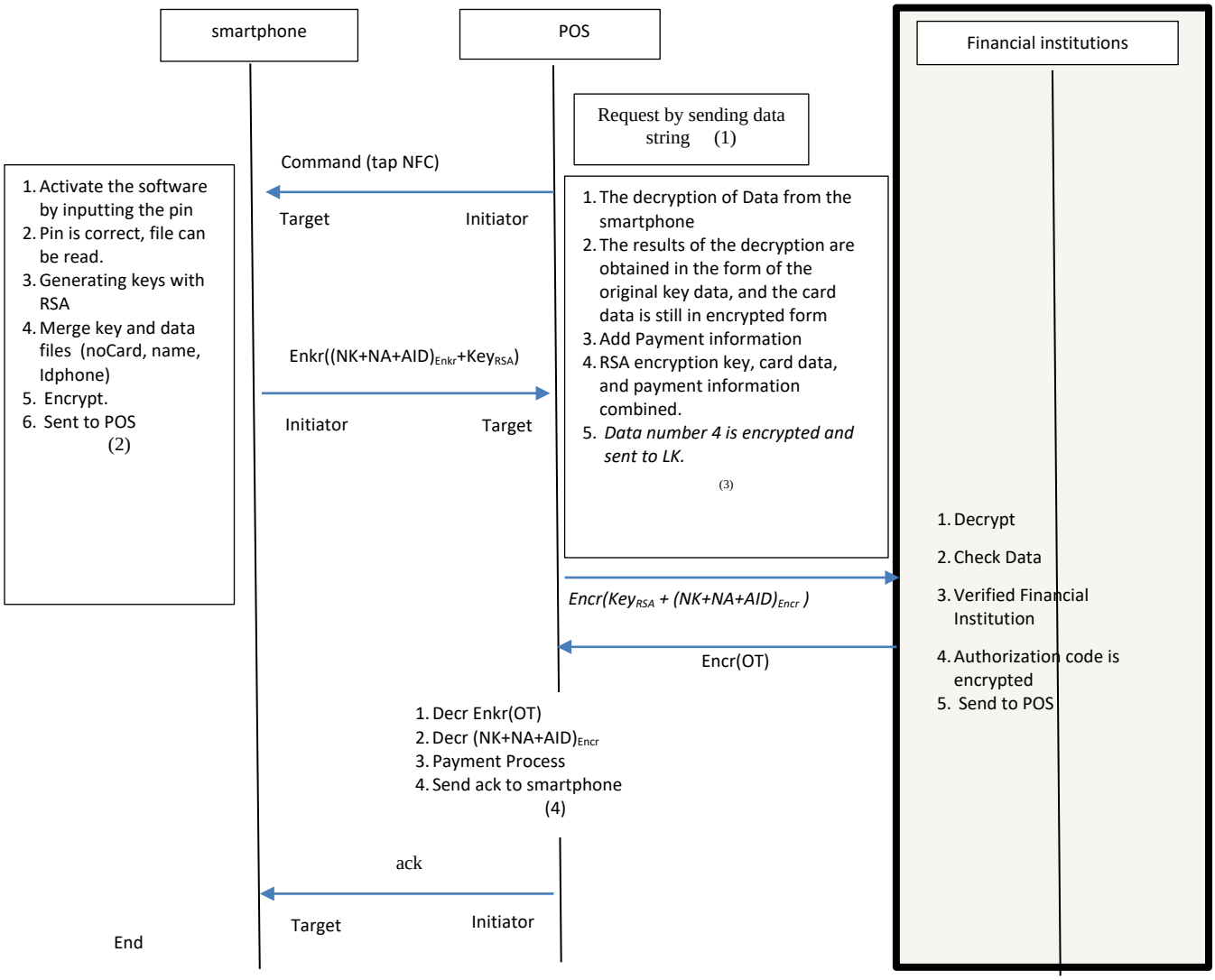


Fig. 2. NFC-HCE Transaction Model without Internet Connection between Smartphone and POS.

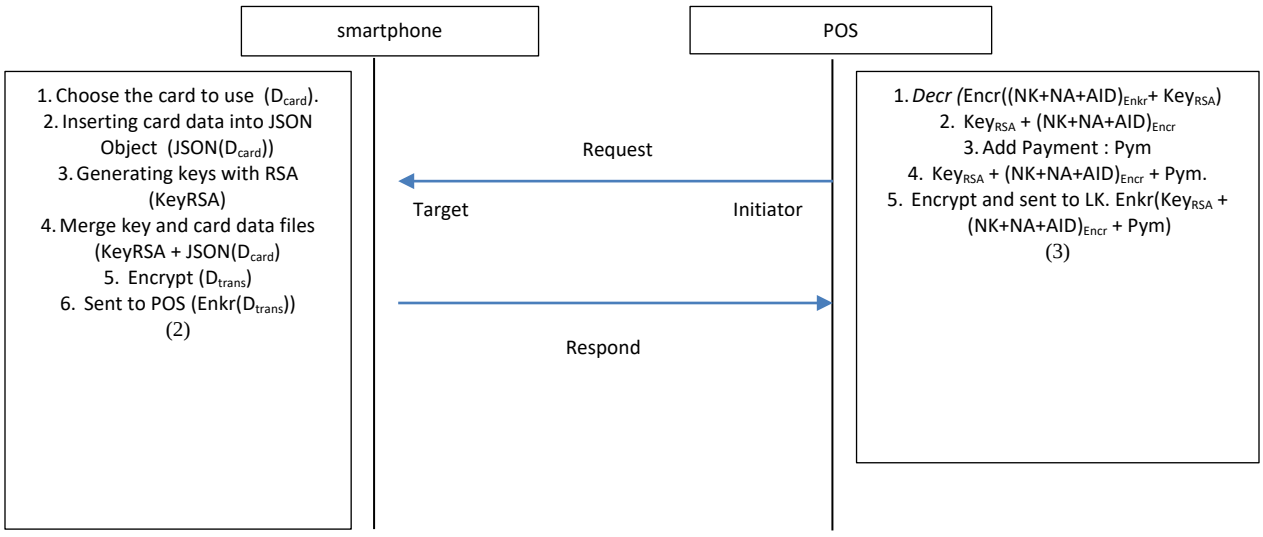


Fig. 3. Transaction Model Steps 2 and 3.

The data sent from the smartphone to the POS will be encrypted and encapsulated first. After the data reaches the POS, the data is parsed to get the original data. Encryption is done using the AES cryptographic algorithm. The algorithm chosen is simple, lightweight, and has modifiable parameters. It is because computing is done on smartphones that have limited memory capacity. At the same time, the parameter options can be modified to be used to add security by changing the parameter value every time there is a new transaction. Until this stage, the transaction model is complete. The model created is a model for normal transactions. Prevention of relay attacks is done by encrypting data when it is stored and transmitted between devices.

IV. Implementation and Evaluation

The proposed transaction model was tested on a smartphone with NFC facilities and the android operating system. The parameters tested were execution time and data randomness analysis using the monobit test.

The model was tested on a smartphone with differences in the android version and memory size. This test ensures that the model runs well and can improve security. Trials were also carried out on different smartphone conditions to determine the model's performance in different situations.

The data used for the trial is dummy data which describes the card user data that is widely used today. This data is inputted into APL-SE at the initialization stage, then used for transactions in this trial.

This encryption time is calculated from the beginning of the encryption process until the data is successfully encrypted. Based on the graphs in Fig. 4 and Fig. 5, we find that the time it takes is far below one second; the average time based on the test results is 1.074 milliseconds. This time is very short compared to the time it takes for the attacker to retrieve and translate the encrypted data, which takes more than one millisecond [25].

The entropy value shown in Fig. 6 and the P value shown in Fig. 7 also indicates that the data encryption results are declared random, based on Shannon Theory.

The communication made by the smartphone and the POS at the time of the transaction is the exchange of credential data needed to declare the transaction successful and correct. In this trial, the smartphone used by the user is a smartphone with the Android operating system, and the one used as a POS is an android smartphone.

Data exchange is carried out using the NFC-HCE ecosystem facilities on mobile devices. Fig. 8 shows the process of exchanging payment amount data from POS to the smartphone.

Data can be sent by POS if there is a request from a smartphone. For this reason, at this time, the smartphone is conditioned as the initiator and the POS as the target. In the beginning, the data sent from the POS to the smartphone is entered into the NFC POS card in the form of a shared preference. When the smartphone taps into the POS, the smartphone sends a request, namely the APDU command. POS receives this APDU command, then processes the APDU command. Payment data and select AID commands are combined and sent to the smartphone. The smartphone will receive the data, then retrieve the payment data. The first request-response process between smartphone (initiator) and POS (target) is finished here.

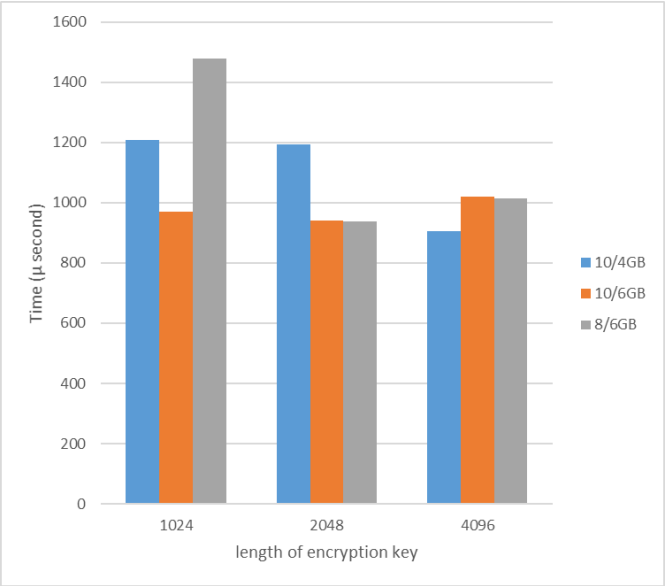


Fig. 4. Data Encryption Time on a Smartphone.

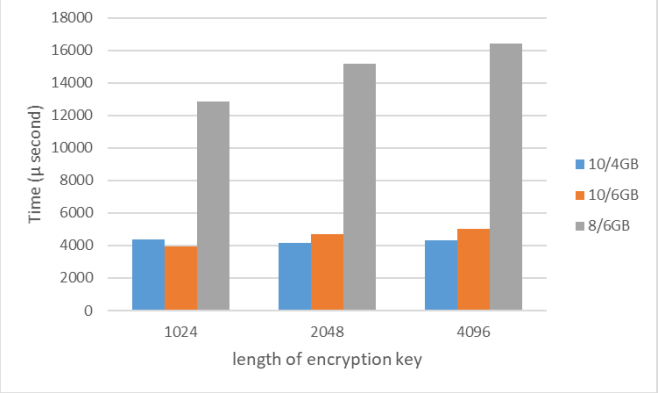


Fig. 5. Time to Send Customer Data to POS via NFC-HCE.

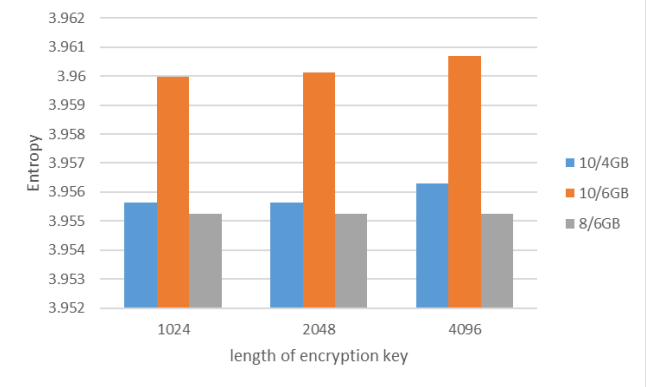


Fig. 6. Entropy Value of Data Encryption Results.

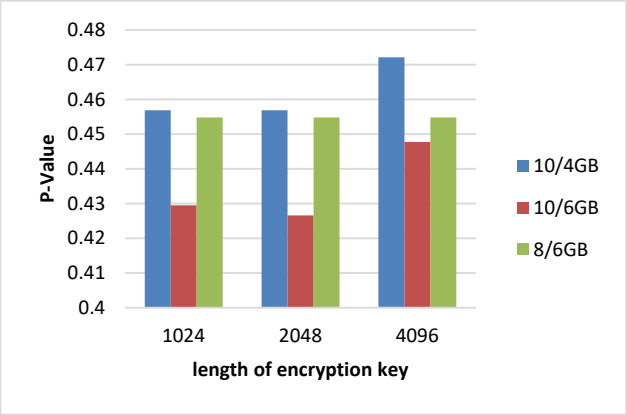


Fig. 7. P Value of Data Encryption Results.

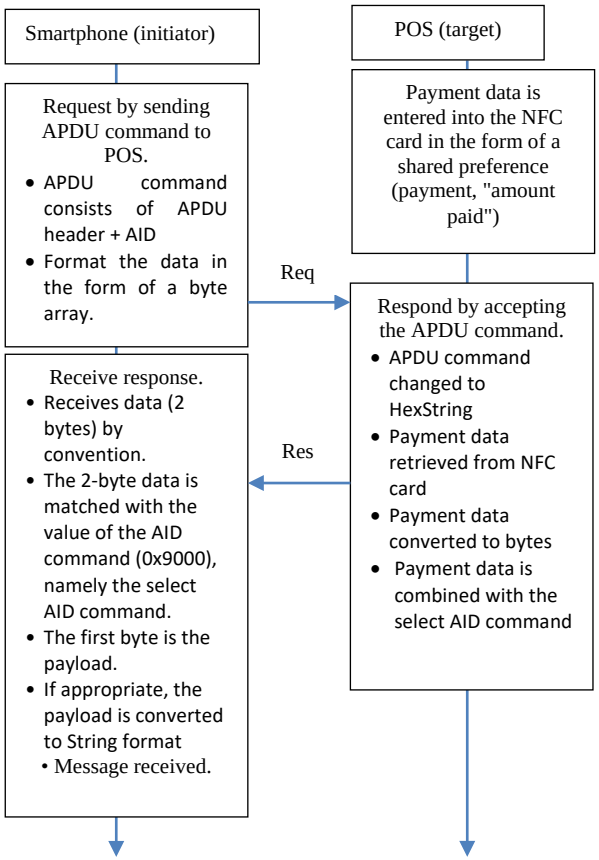


Fig. 8. Request-Respond when Sending Payment Data from POS to Smartphone.

The position of the initiator and target changes, POS as the initiator will make the request, and the smartphone as the target will respond. Fig. 9 shows this process. This process is similar to the process in Fig. 5. The data sent from the smartphone is the card data that will be used for payments. Card data is RSA encrypted in a JSON Object and sent in a shared preference format. When the card data reaches the POS, it is combined with the payment data, and the POS sends it to the server for verification.

Trials were also carried out by simulating smartphones with various positions during transactions. The model is tested using several smartphones to simulate a situation where there is another NFC within reach of the NFC reading area (POS or user) during a transaction. The test shows that the NFC that is read is the NFC that is closer to the NFC reader (reader/initiator). Because the physical situation that must be done is a situation where communication occurs with an NFC tap, then the two NFCs that will communicate are at a very close distance and do not allow other smartphones to occupy a closer position. This physical situation prevents NFC reading errors with unexpected devices.

NFC on smartphones, both POS and users, can be read at a distance of 0-5 cm. The NFC is not readable if their distance is above that distance. So, in that situation, the attack could be prevented because:

1) The data sent is in encrypted form.

2) The distance between smartphones must be close, and the NFC smartphone must be facing, meaning that the back of the smartphone must also be facing the back of the other.

3) The distance between smartphones is not more than 5 cm.

The second situation tested is when the user and the POS communicate, there is another NFC on the side of the user's NFC opposite the POS. When this happens, and the user's NFC is in reading mode (target), the user's NFC will be read by the POS because the NFC position is on the back side of the smartphone, which is facing the POS when communicating. Likewise, if the user's NFC is in reading mode (initiator), the user's NFC will read the NFC POS because the NFC position is on the back side of the smartphone facing the POS. This situation was tested by varying the distance between smartphones. 0-5 cm distance and NFC facing each other (meaning the back of the smartphone is facing each other) can communicate.

The third situation that is tested is if the POS and the user are going to make a transaction, and it turns out that another smartphone is in a position between the two devices. The first possibility is that the POS will read the NFC of another smartphone. Thus, communication between the POS and the user does not occur. If there were such a situation, the possible situation would be as follows:

1) If another smartphone has APL-SE, the transaction can occur, but the transaction is between the POS and another smartphone. In the transaction, authentication is carried out on the device's pin and id. If authentication is performed and another smartphone is identified, then the payment is made by the other smartphone and the accompanying identity.

2) If the other smartphone does not have APL-SE, there will be no transaction because the transaction will only occur if the two devices are connected as POS and user.

The second possibility is that the POS will read the user's smartphone because the NFC is the NFC of the POS and the user. So even though there are closer smartphones, because the NFC smartphones are not facing each other (there is NFC which is more in a straight line position), there is no NFC communication.

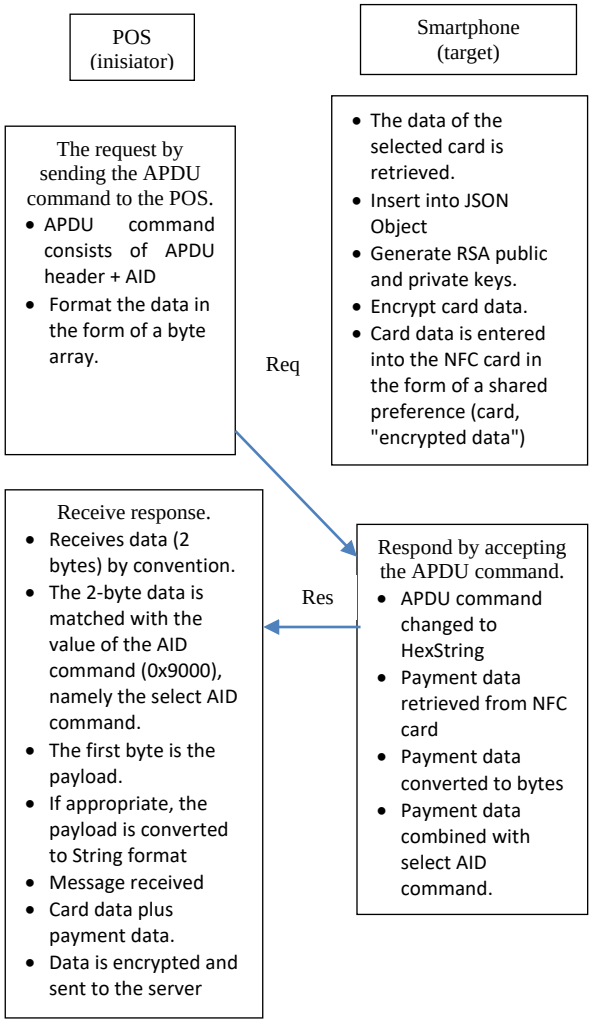


Fig. 9. Request-Response Sending Card Data from Smartphone to POS.

Mutual authentication between buyer and seller is as follows. The initialization model in this study ensures the client by registering client data, card data, and device data. The transaction model ensures that data is safe when sent and received back by the client device because the data is encrypted by each item and encrypted again as a single record before being sent. Smartphone users ensure POS for receiving payment data; POS ensure users receive payment card data. POS ensures financial institutions because the data is verified, and financial institutions ensure POS because the card data sent is correct and in accordance with smartphone owners.

Analysis for security against attackers is carried out as follows. The first attack is an attack that can take data at the time of a transaction, then use it at another time to ensure that the transaction is carried out by the right person (the real account owner). Several processes, namely prevent this attack:

1) Credential data stored on the smartphone is encrypted.

2) When the POS requests the customer's smartphone, the smartphone responds by sending encrypted personal data. The process of sending customer data to the POS can only be done when the customer enters a pin into the system.

3) Sending data from the POS to the Financial Institution for verification is done after the encrypted data from the smartphone is combined with the payment data from the POS, then encrypted.

4) Data originating from smartphones includes customer data and smartphone data. The two data are data pairs registered during the initialization process.

The second attack is an attack that can retrieve and analyze data and how the data can remain safe even though the attacker successfully retrieves the data. Several processes namely prevented the attack:

1) Customer data stored on the smartphone is in an encrypted state. And the results of the analysis of the encrypted data state that the data is random.

2) The encryption process and data transmission process takes place very quickly, far below the time required by the attacker to retrieve and analyze data.

3) Every time a transaction is made, a transaction key is generated, which will be different for other transactions.

The overall process created for this transaction model ensures that transactions are carried out safely and can prevent attacks, especially attacks by retrieving data that is being transmitted and attacks on transactions that use data that is already owned by the attacker.

Authentication in this research is kept simple and doesn't take many steps, but it can ensure that the registered account is correct. So, the account stored on the smartphone is an authenticated account, including the authentication of the smartphone device. Meanwhile, on the level of security, this study uses several levels of security. The first level with an encryption key, the second level encrypts every field of data, the third level encrypts all account data, and the fourth level converts the format to Base64 format.

The condition that requires NFC communication at a short distance and having to face each other on the back of the smartphone is one of the advantages so that the position of an attacker using a smartphone will be difficult. Meanwhile, if there is an attacker who uses another device but can access the data sent in APDU format, then there is a possibility that a third party can retrieve the data. But the format of the data sent does not allow the data to be interpreted in a fast time.

The model made in this study is quite reliable because the use of NFC will continue to grow in the future. The convenience of NFC with just a tap is an advantage, and the tendency of people to do and get more practical things from time to time is increasing.

The implementation of the initialization and transaction model is simple. The model is simply applied to the payment system at the POS, and financial institutions verify their customer account data, and smartphone owners download payment applications. The condition for this model to be implemented is that there is an agreement between POS and financial institutions, and there are customers who use applications that implement this model.

V. Conclusion

The APL-SE transaction model was created to improve the security of payment transactions using NFC-enabled mobile in the NFC-HCE ecosystem. The NFC-HCE ecosystem will be increasingly used because of its practicality in not needing to use a SIM as a place to store secure elements. Thus, financial institutions do not need to rent space from mobile operators. The test results show that the processing time is short and the encrypted data is random, thus increasing security.

This study has not discussed and tested data stored on smartphones when not in use. Data stored on smartphones has many security variables, such as user negligence, and lost or damaged data.

Acknowledgment

The study was supported by "Indonesia Digital Technology University."

References

[1] F. Fainusa, R. Nurcahyo, and M. Dachyar, "Conceptual Framework for Digital Wallet User Satisfaction," ICETAS 2019 - 2019 6th IEEE Int. Conf. Eng. Technol. Appl. Sci., pp. 2019–2022, 2019, doi: 10.1109/ICETAS48360.2019.9117285.

[2] Y. U. Chandra, Ernawaty, and Suryanto, "Bank vs telecommunication E-Wallet : System analysis, purchase, and payment method of GO-mobile CIMB Niaga and T-Cash Telkomsel," Proc. 2017 Int. Conf. Inf. Manag. Technol. ICIMTech 2017, vol. 2018-Janua, no. May, pp. 165–170, 2018, doi: 10.1109/ICIMTech.2017.8273531.

[3] Abdullah Almuhammadi, "An Overview of Mobile Payment, Fintech, and Digital Wallet in Saudi Arabia," 2020 7th Int. Conf. Comput. Sustain. Glob. Dev., pp. 271–278, 2020, doi: 10.23919/INDIACom49435.2020.9083726.

[4] R. M. N. D. Ranasinghe, "RFID / NFC Device with Embedded Fingerprint Authentication System," pp. 21–24, 2017, doi: 978-1-5???-??97-7/1??$31.00.

[5] O. S. Okpara and G. Bekaroo, "Cam-Wallet: Fingerprint-based authentication in M-wallets using embedded cameras," Conf. Proc. - 2017 17th IEEE Int. Conf. Environ. Electr. Eng. 2017 1st IEEE Ind. Commer. Power Syst. Eur. EEEIC / I CPS Eur. 2017, pp. 1–5, 2017, doi: 10.1109/EEEIC.2017.7977654.

[6] J. Zhao and X. Y. Li, "SCsec: A Secure near Field Communication System via Screen Camera Communication," IEEE Trans. Mob. Comput., vol. 19, no. 8, pp. 1943–1955, 2020, doi: 10.1109/TMC.2019.2913412.

[7] P. Pourghomi, P. E. Abi-char, and G. Ghinea, "Towards a mobile payment market: A Comparative Analysis of Host Card Emulation and Secure Element," Int. J. Comput. Sci. Inf. Secur., vol. 13, no. 12, pp. 156–164, 2015.

[8] L. N. Harnaningrum, A. Ashari, and A. E. Putra, "SECURE INITIALIZATION MODEL IMPROVEMENT for NFC-HCE SECURITY in MOBILE PAYMENT SYSTEM," J. Theor. Appl. Inf. Technol., vol. 99, no. 24, pp. 6139–6151, 2021.

[9] M. Badra and R. B. Badra, "A Lightweight Security Protocol for NFC-based Mobile Payments," Procedia Comput. Sci., vol. 83, no. Ant, pp. 705–711, 2016, doi: 10.1016/j.procs.2016.04.156.

[10] S. Nashwan, "Secure Authentication Protocol for Mobile Payment," Int. J. Comput. Sci. Netw. Secur., vol. 17, no. 8, pp. 256–263, 2017, doi: 10.26599/tst.2018.9010031.

[11] M. A. Crossman and H. Liu, "Two-factor authentication through near field communication," 2016 IEEE Symp. Technol. Homel. Secur. HST 2016, 2016, doi: 10.1109/THS.2016.7568941.

[12] A. Munch-Ellingsen, R. Karlsen, A. Andersen, and S. Akselsen, "Two-factor authentication for android host card emulated contactless cards," 2015 1st Conf. Mob. Secur. Serv. MOBISECSERV 2015, 2015, doi: 10.1109/MOBISECSERV.2015.7072874.

[13] M. H. Özcanhan, G. Dalkiliç, and S. Utku, "Cryptographically supported NFC tags in medication for better inpatient safety patient facing systems," J. Med. Syst., vol. 38, no. 8, 2014, doi: 10.1007/s10916-014-0061-x.

[14] F. Dang et al., "Large-scale invisible attack on AFC systems with NFC-equipped smartphones," Proc. - IEEE INFOCOM, 2017, doi: 10.1109/INFOCOM.2017.8057219.

[15] M. Sujithra, G. Padmavathi, and S. Narayanan, "Mobile device data security: A cryptographic approach by outsourcing mobile data to cloud," Procedia Comput. Sci., vol. 47, no. C, pp. 480–485, 2015, doi: 10.1016/j.procs.2015.03.232.

[16] M. Al-fayoumi and S. Nashwan, "Performance Analysis of SAP-NFC Protocol," Int. J. Commun. Networks Inf. Secur., vol. 10 No 1, no. April, p. 125, 2018.

[17] M. Alattar and M. Achemlal, "Host-based card emulation: Development, security, and ecosystem impact analysis," Proc. - 16th IEEE Int. Conf. High Perform. Comput. Commun. HPCC 2014, 11th IEEE Int. Conf. Embed. Softw. Syst. ICESS 2014 6th Int. Symp. Cybersp. Saf. Secur., pp. 506–509, 2014, doi: 10.1109/HPCC.2014.85.

[18] A. Asaduzzaman, S. Mazumder, and S. Salinas, "A Security-Aware Near Field Communication Architecture," 2017 Int. Conf. Networking, Syst. Secur., no. January, 2017.

[19] N. Alzahrani, "Securing Pharmaceutical and High-Value Products Against Tag Reapplication Attacks Using NFC Tags," 2016 IEEE Int. Conf. Smart Comput., 2016, doi: 10.1109/SMARTCOMP.2016.7501715.

[20] O. Wenxing, W. Lei, Z. Yu, and Y. Changhong, "Research on Anti-eavesdropping Communication Mechanism for NFC," Proc. - 2015 7th Int. Conf. Meas. Technol. Mechatronics Autom. ICMTMA 2015, pp. 839–841, 2015, doi: 10.1109/ICMTMA.2015.206.

[21] K. Fan, P. Song, and Y. Yang, "ULMAP : Ultralightweight NFC Mutual Authentication Protocol with Pseudonyms in the Tag for IoT in 5G," Mob. Inf. Syst., vol. 2017, no. April, 2017.

[22] N. El Madhoun, E. Bertin, and G. Pujolle, "For Small Merchants: A Secure Smartphone-Based Architecture to Process and Accept NFC Payments," Proc. - 17th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. 12th IEEE Int. Conf. Big Data Sci. Eng. Trust. 2018, pp. 403–411, 2018, doi: 10.1109/TrustCom/BigDataSE.2018.00067.

[23] F. Dang et al., "Pricing Data Tampering in Automated Fare Collection with NFC-Equipped Smartphones," IEEE Trans. Mob. Comput., vol. 18, no. 5, pp. 1159–1173, 2019, doi: 10.1109/TMC.2018.2853114.

[24] A. Al-Haj and M. A. Al-Tameemi, "Providing security for NFC-based payment systems using a management authentication server," 2018 4th Int. Conf. Inf. Manag. ICIM 2018, pp. 184–187, 2018, doi: 10.1109/INFOMAN.2018.8392832.

[25] N. El Madhoun, E. Bertin, and G. Pujolle, "An overview of the EMV protocol and its security vulnerabilities," 2018 4th Int. Conf. Mob. Secur. Serv. MOBISECSERV 2018, vol. 2018-Febru, no. February, pp. 1–5, 2018, doi: 10.1109/MOBISECSERV.2018.8311444.

[26] D. Mahansaria and U. K. Roy, "Secure authentication for ATM transactions using NFC technology," Proc. - Int. Carnahan Conf. Secur. Technol., vol. 2019-Octob, pp. 1–5, 2019, doi: 10.1109/CCST.2019.8888427.

Observation of Imbalance Tracer Study Data for Graduates Employability Prediction in Indonesia

Ferian Fauzi Abdulloh, Majid Rahardi, Afrig Aminuddin, Sharazita Dyah Anggita, Arfan Yoga Aji Nugraha
Computer Science Faculty, University of AMIKOM Yogyakarta
Yogyakarta, Indonesia

***Abstract*—Tracer Study is a mandatory aspect of accreditation assessment in Indonesia. The Indonesian Ministry of Education requires all Indonesia Universities to anually report graduate tracer study reports to the government. Tracer study is also needed by the University in evaluating the success of learning that has been applied to the curriculum. One of the things that need to be evaluated is the level of absorption of graduates into the working industry, so a machine learning model is needed to assist the University Officials in evaluating and understanding the character of its graduates, so that it can help determine curriculum policies. In this research, the researcher focuses on making a reliable machine learning model with a tracer study dataset format that has been determined by the Government of Indonesia. The dataset was obtained from the tracer study of Amikom University. In this study, SVM will be tested with several variants of the algorithm to handle imbalanced data. The study compared SMOTE, SMOTE-ENN, and SMOTE-Tomek combined with SVM to detect the employability of graduates. The test was carried out with K-Fold Cross Validation, with the highest accuracy and precision results produced by SMOTE-ENN SVM model by value of 0.96 and 0.89.**

Keywords*—Tracer study; support vector machine; synthetic minority oversampling technique; SMOTE; employability

I. Introduction

A decent University can be seen from the level of absorption of its graduates in working world, thus many universities are trying to improve the quality of their graduates [1], [2]. That is the reason why the Indonesian Ministry of Education requires all Universities to always report the results of tracer study anually for measuring University graduates employability. Tracer study is also a requirement for higher education accreditation set by the National Accreditation Board for Higher Education (BAN-PT) [3], [4].

Currently we live surrounded by data, data circulating around us can be collected and processed to produce new knowledge [5], including tracer study data. These data can be collected and processed to improve the quality of human resources and curriculum that can increase the absorption of university graduates in industries.[1], [6].

One of the machine learning models that have been widely used to meet these needs is classification [7], [8]. Using classification algorithm we can predict whether an alumni has the possibility of being absorbed in a job quickly or not [9].

There are many classification algorithms that are popularly used, one of which is the Support Vector Machine, from previous research the SVM algorithm is very well used to predict the employability of graduates [10], but basically the final result of an algorithm does not only depend on the quality of the algorithm used but also on the quality of the dataset applied to the algorithm, one of the criteria to get a reliable machine learning model is that the dataset must be balanced, to balance the dataset there are 2 methods, namely oversampling and undersampling, one of the oversampling algorithms that can be used is SMOTE, SMOTE itself has several variants, namely SMOTE, SMOTE ENN, and SMOTE Tomek[11], [12].

This study aims to find out the best method for predicting the employability of higher education alumni using the Amikom University tracer study dataset with attributes and formats determined by the Indonesian Ministry of Education which can be accessed on the web http://tracerstudy.kemdikbud.go.id/ frontend/.

II. Literature Review

A. Classification

Classification is a type of machine learning algorithm where the computer will automatically predict the class of a data from the input data given [7]. Several classification algorithms commonly used for tracer studies include Naive Bayes, Neural Network, SVM, Logistic Regression, etc [9], [13], [14]. In previous works, Tracer Study Data in Indonesia was analyzed using those classification algorithms, without using SMOTE or another imbalanced data handler model.

B. Balance Data

Balanced dataset is data in which the comparison of each data in a class is balanced, the data in which each class has a significantly different amount, the dataset is called imbalance. Unbalanced classes are a common problem in machine learning classification where there is a disproportionate ratio in each class. Class imbalances can be found in various fields , moreover in tracer study case. Classes that have more data are often called majority classes and classes that have less data are called minority classes[15]–[17].

C. Support Vector Machine

The Support Vector Machine algorithm is one of the algorithms included in the Supervised Learning category, which means that the data used for machine learning is data that has a previous label[18], [19]. So that in the decision-making process, the machine will categorize the testing data into labels that are in accordance with its characteristics.

Support Vector Machine is one of the machine learning algorithms that can be used for classification, where this algorithm will generate the best hyperplane where this hyperplane will separate the classes in the dataset [20], [21].

$$w.\ x - b = 0 \quad (1)$$

where:

w = Weight Vector

x = Input Vector

b = Bias

D. SMOTE

SMOTE is one of the algorithms that can be used to balance a dataset, using an oversampling approach, in which this algorithm will generate synthesis data from the minority class so that the minority class has the same amount of data as the majority class [15], [22].This synthetic data is obtained based on the value of k-neighbours from minority data.

$$\Delta(A,B) = W_A W_B \sum_{i=1}^{N} \delta(v1.v2)^r \quad (2)$$

$\Delta(A,B)$: observed distance between A & B
$W_A W_B$: observed weight
N: amounts of predictor variables
r: value of 1 (Manhattan distance) or 2 (Euclidean dist)
$\delta(v1.v2)^r$: the distance between observations A and B for each explanatory variable, with the formula;

$$\delta(v1.v2) = \sum_{i=1}^{n} \left| \frac{c_{1i}}{c_1} - \frac{c_{2i}}{c_2} \right| \quad (3)$$

$\delta(v1.v2)$: the distance between observations A and B which is included in the i variable

c_{1i}: the number of the 1st category which is included in the i-th explanatory variable category

c_{2i}: the number of the 2nd category which is included in the i-th explanatory variable category

c_1: number of category 1

c_2: number of category 2

n: the number of categories in the i-th explanatory variable

k: Constant

In this study, researchers will compare three variants of the SMOTE algorithm, namely, SMOTE, SMOTE ENN and SMOTE Tomek. SMOTE Tomek uses a combination of the SMOTE algorithm which is a balancing algorithm with an oversampling approach combined with ENN and Tomek which is an undersampling algorithm, where ENN and Tomek function to delete synthetic data that has similarities to the majority data so that data balance is obtained where each data class has a clear difference [11], [23].

III. RESEARCH METHOD

The dataset used in this study is data obtained from questionnaires filled out by alumni of Amikom University in 2018. The questionnaires that have been distributed are then filled out by (many) respondents and stored in csv form. The process can be seen in the Fig 1.

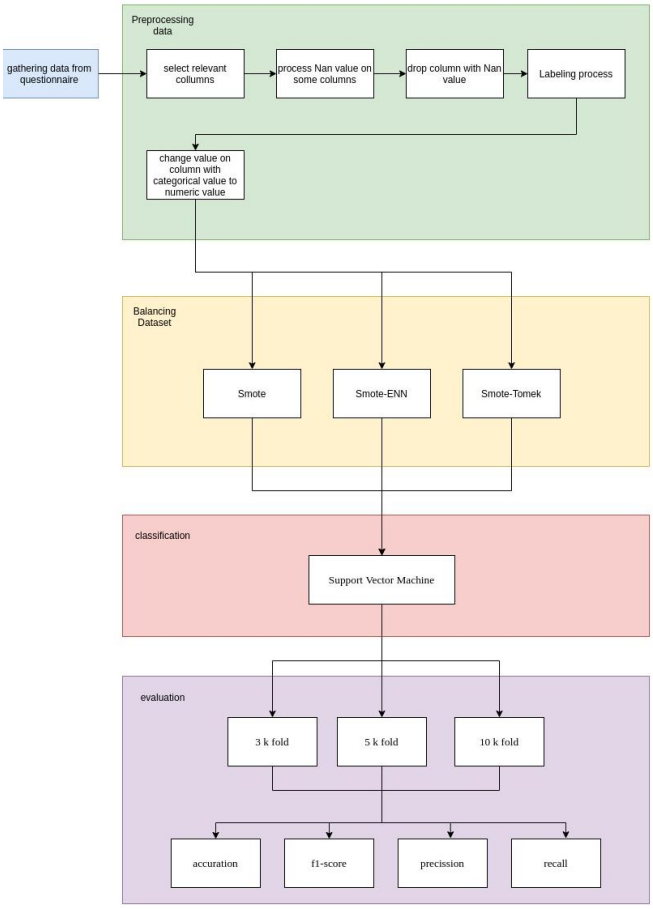


Fig. 1. Research Overview

A. Selection of Attributes and Collection of Survey Results

The first stage of this research is to collect the results of the questionnaire; which later the results from this questionnaire will be presented in csv form so that thereafter it can be processed using a predetermined model. There are 145 collumns consists of their hardskill level after graduate, sex, how long they study in college, when they start to search jobs, and many more, including the label (alumni employability). All of the atributes can be accessed at http://tracerstudy.kemdikbud.go.id/ frontend/.

B. Labeling Data

Data labeling is done by taking each respondent's answer to the question "How long did it take you to get your job after graduation?" In this research, based on that question, labels are divided into three classes. If a student gets a job before graduating from University, then the data will be labeled as "1". If the student gets a job three months or less after they graduate from University then it will be labeled "2". If the student takes more than 3 months to get a job get a job after graduation it will be labeled "3".

C. Data Preprocessing

In this process, preprocessing of data is carried out by converting data labeled string into integer form and also filling empty values in all existing columns with zero values, and deleting values with remaining null data. This have to be done to avoid anomalies in the mathematical modeling.

D. Data Balancing

In practice, classification requires balanced data, balanced data is data where each label has the same amount, if each label has a significantly different amount then the dataset is called imbalanced. Class that has more data is the majority class and the class that has less data is called the minority class [24].

In this study, to overcome the imbalanced data, SMOTE algorithm is used, SMOTE is an algorithm that is useful for balancing the amount of data with an oversampling approach, the SMOTE algorithm will create synthesis data obtained based on the value of k-neighbours from minority data [25].

E. Classification

After the data balancing process, the classification process is carried out with the Support Vector Machine algorithm.

F. Testing

The model testing process uses the K-Fold Cross Validation algorithm with Folds determined to be 3, 5, and 10 Folds. This is done so that the test is more valid and vary [26].

IV. Results

In this study, we will classify the normalized tracer study dataset. After collecting and normalizing the dataset , the dataset will be divided into three classes based on when the alumni got a job, the first class will contain data on alumni who got a job before graduating, less than or three months after graduation, and more than three months after graduating. Fig. 2 showed us the amount of data that has imbalance class. Fig. 3 showed that the amount of dataset significantly altered in every observation using different types of SMOTE.

There are three models of balancing algorithm that will be compared, those are SMOTE, SMOTE ENN and SMOTE Tomek algorithms when applied to the support vector machine classification algorithm. The best model will be calculated based on the average value of f1, accuracy, precision, and recall.

The SMOTE algorithm is a data balancing algorithm with an oversampling approach where the number of minority classes will be increased to balance the majority class. Fig. 4-6 show the dataset after being applied to SMOTE, SMOTE ENN and SMOTE Tomek algorithms

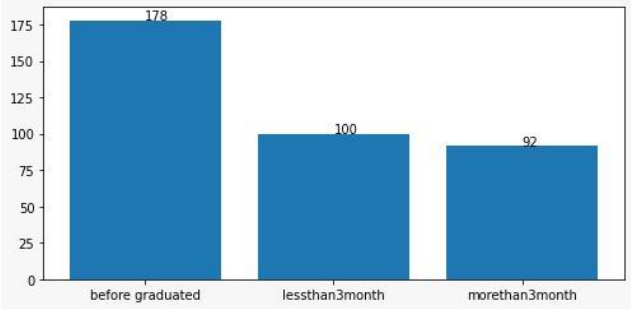


Fig. 2. Dataset before Balancing

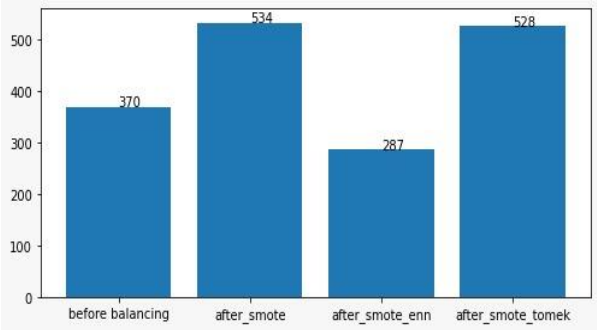


Fig. 3. All Dataset Amounts before and after Balancing

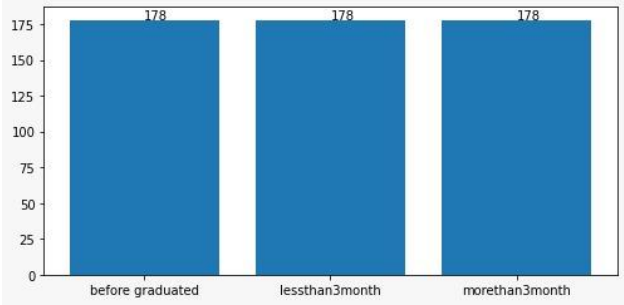


Fig. 4. Dataset after SMOTE

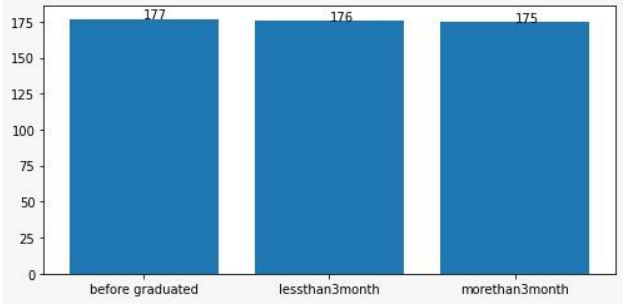


Fig. 5. Dataset after SMOTE-Tomek

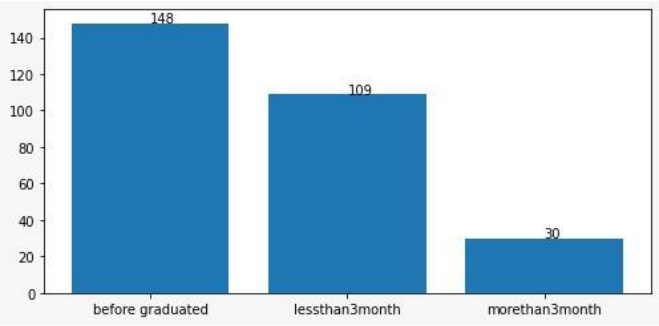


Fig. 6. Dataset after SMOTE-ENN

After the dataset are being processed by SMOTE and SMOTE-TOMEK algorithms, it produces classes that have balanced amount of data. But it did not happen in the SMOTE ENN algorithm, SMOTE ENN created a more normal dataset, this is because when data has an absolute balance, sometimes it may result in overfitting.[12]

Furthermore, after getting the data that we have balanced, the data will be applied to the Support vector machine classification algorithm and for model level measurements, cross fold validation measurements will be used with 3, 5 and 10 fold values for accuracy, f1 score, recall and precision for each model.

TABLE I. RESULT OF ACCURACY & F1 SCORE 3-FOLD

balancing	accuracy			avg	F1 score			avg
-	0.86	0.77	0.76	0.79	0.84	0.73	0.73	0.76
smote	0.78	0.86	0.96	0.86	0.77	0.86	0.96	0.86
smote-enn	0.98	0.94	0.94	0.95	0.97	0.86	0.86	0.89
smote-Tomek	0.86	0.91	0.95	0.90	0.86	0.91	0.95	0.90

TABLE II. RESULT OF RECALL & PRECISION 3-FOLD

balancing	precision			avg	recall			avg
-	0.86	0.82	0.82	0.83	0.83	0.72	0.70	0.75
smote	0.78	0.88	0.96	0.87	0.78	0.86	0.96	0.86
smote-enn	0.98	0.95	0.95	0.96	0.96	0.82	0.82	0.86
smote-Tomek	0.87	0.92	0.96	0.91	0.86	0.91	0.95	0.90

Shown in Table I and II, the experiment is done by using three fold cross validation to test the f1 score, accuracy, precision, and recall from SVM with SMOTE, SMOTE-TOMEK, and SMOTE_ENN and the results obtained that this research scenario has an average f1 accuracy result. score, precision and recall using SVM alone are 0.79.0.76, 0.83, 0.75 and after data balancing, the f1 score, precision and recall are respectively as follows

Smote : 0.86, 0.86, 0.87, 0.86

smote -enn : 0.95, 0.89 ,0.96, 0.86

Smote-tomek : 0.90, 0.90, 0.91, 0.90

These results indicate that the Three-Fold SMOTE, SMOTE-Tomek, and SMOTE-ENN validations are proven to be able to increase the accuracy value of SVM itself, with the highest average value generated by SVM Smote-ENN.

TABLE III. RESULT OF ACCURACY 5-FOLD

	balancing	accuration					avg
1	-	0.93	0.80	0.78	0.86	0.85	0.84
2	smote	0.86	0.75	0.89	0.94	0.99	0.88
3	smote-enn	0.93	0.98	0.98	0.96	0.91	0.95
4	smote-Tomek	0.97	0.82	0.91	0.94	0.99	0.92

TABLE IV. RESULT OF F1 SCORE 5-FOLD

	balancing	f1-score					avg
1	-	0.93	0.76	0.74	0.85	0.84	0.88
2	smote	0.86	0.75	0.89	0.94	0.99	0.82
3	smote-enn	0.89	0.96	0.99	0.92	0.79	0.91
4	smote-Tomek	0.97	0.82	0.90	0.94	0.99	0.92

TABLE V. RESULT OF RECALL 5-FOLD

	balancing	recall					avg
1	-	0.91	0.75	0.73	0.84	0.81	0.80
2	smote	0.86	0.75	0.89	0.94	0.99	0.88
3	smote-enn	0.86	0.94	0.98	0.89	0.76	0.88
4	smote-Tomek	0.97	0.82	0.91	0.94	0.99	0.92

TABLE VI. RESULT OF PRECISION 5 FOLD

	balancing	precision					avg
1	-	0.96	0.79	0.84	0.87	0.92	0.87
2	smote	0.86	0.78	0.91	0.95	0.99	0.89
3	smote-enn	0.95	0.99	0.99	0.97	0.93	0.96
4	smote-Tomek	0.97	0.84	0.92	0.95	0.99	0.93

In the test scenario using five cross fold validation that are shwon at Table III, IV, V and VI, the average results of the f1 score accuracy, precision and recall are 0.84, 0.88, 0.87, 0.80 after data balancing the f1 score accuracy, precision and recall values are equal to

Smote: 0.88, 0.82, 0.89, 0.88

Smote -enn: 0.95, 0.91 ,0.96, 0.88

Smote-tomek: 0.92, 0.92, 0.93, 0.92

then it can be seen from the data that the values of accuracy, precision, recall and f1 are close to perfect which indicates an overfitting, this is triggered by the distribution of test data that is less than the previous experiment.

TABLE VII. RESULT OF NO-SMOTE & SMOTE 10-FOLD

Sub set	**Support vector machine**							
	Without balancing data				**smote**			
	acuracy	precision	recall	f1	accuracy	precision	recall	F1
1	0.86	0.91	0.82	0.84	0.89	0.89	0.89	0.89
2	0.95	0.97	0.93	0.95	0.85	0.86	0.85	0.85
3	0.86	0.90	0.83	0.85	0.76	0.79	0.76	0.76
4	0.76	0.75	0.71	0.72	0.81	0.84	0.81	0.82
5	0.70	0.73	0.62	0.62	0.87	0.91	0.87	0.87
6	0.81	0.88	0.76	0.77	0.94	0.95	0.94	0.94
7	0.95	0.95	0.93	0.94	0.98	0.98	0.98	0.98
8	0.78	0.78	0.73	0.75	0.94	0.95	0.94	0.94
9	0.86	0.92	0.83	0.85	1.00	1.00	1.00	1.00
10	0.86	0.92	0.83	0.86	0.98	0.98	0.98	0.98
avg	**0.84**	**0.87**	**0.80**	**0.81**	**0.9**	**0.91**	**0.9**	**0.9**

TABLE VIII. RESULT OF SMOTE-ENN & SMOTE-TOMEK 10-FOLD

Sub set	**Support vector machine**							
	Smote ENN				**Smote tomek**			
	accuracy	**precission**	**recall**	**f1**	**accuracy**	**precission**	**recall**	**F1**
1	0.97	0.97	0.89	0.92	1.00	1.00	1.00	1.00
2	0.93	0.95	0.86	0.89	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	0.81	0.85	0.81	0.80
4	0.97	0.97	0.89	0.92	0.87	0.88	0.87	0.87
5	0.97	0.98	0.97	0.97	0.92	0.93	0.92	0.92
6	1.00	1.00	1.00	1.00	0.94	0.95	0.94	0.94
7	0.97	0.97	0.89	0.92	1.00	1.00	1.00	1.00
8	0.96	0.97	0.89	0.92	0.91	0.91	0.91	0.91
9	0.93	0.95	0.78	0.81	1.00	1.00	1.00	1.00
10	0.89	0.92	0.74	0.77	0.98	0.98	0.98	0.98
avg	**0.95**	**0.96**	**0.89**	**0.91**	**0.94**	**0.94**	**0.94**	**0.94**

Just like the previous two experiments in the 10 cross fold validation experiment that can be read in Table VII and Table VIII, before the application of balancing the data model, the accuracy value was equal to 0.84, f1 was equal to 0.81, precision was equal to 0.87 and recall is 0.8, then after SMOTE being implemented, there was an increase in the accuracy of the f1 score, precision and recall. The four values increase after data balancing is done. The value of f1 score accuracy, precision and recall is equal to getting the average result

Smotes : 0.90, 0.90, 0.90, 0.91

smooth-enn : 96, 91 ,96, 89

Smote-tomek : 0.94, 0.94, 0.94, 0.94

However, in this experiment, it can be seen that there is an overfitting of the SVM model that uses a data balancing algorithm in several folds which is marked by perfect accuracy in all 3 algorithms. This happens because the test data is only 10% of the entire dataset, it can also be seen in the ENN and Tomek algorithms, cases of overfitting occur more than in the smote algorithm, this is due to the significant difference between the classes in the dataset after the application of the enn and tomek algorithms which is getting worse. enlarge the difference in the data in each class.

V. CONCLUSION

In this study, data balancing algorithms smote, and smote tomek can be used to produce balanced data in terms of the balance ratio formula. Both of these algorithms also produce accuracy, f1 score, precision and recall which are quite significant considering the results presented. However, compared to the SMote-ENN algorithm which produces a poor balance ratio value, the smote tomek and smote algorithms have a lower accuracy value of f1 score, precision and recall. Several fold-cross validation were performed to analyze the data, and found that SMOTE-ENN has the best accuracy in general. In 10-Fold Validation Without SMOTE produced 0.84 in accuracy, using SMOTE it produced 0.9 in accuracy, using SMOTE-Tomek it has 0.94 in accuracy point, and the last one SMOTE-ENN has 0.95 in accuracy.

The SMOTE-ENN-SVM algorithm produces a model with better quality, this can be seen from the accuracy score in each experiment which is higher than other algorithms. In the future, because Tracer Study Data that has many collumns and vary type of data, it would be better to perform feature selection algorithms to select the best feature to be analyzed.

ACKNOWLEDGMENT

This research has been fully supported by grants from the AMIKOM University (#RP-1585824970000).

REFERENCES

[1] A. C. Albina and L. P. Sumagaysay, “Employability tracer study of Information Technology Education graduates from a state university in the Philippines,” Social Sciences & Humanities Open, vol. 2, no. 1, p. 100055, 2020, doi: 10.1016/j.ssaho.2020.100055.

[2] A. F. Hasibuan1, S. M. Silaban2, F. Lubis3, and R. R. Prayogo, “Tracer Study Exploration of Medan State University Graduates,” 2020. [Online]. Available: http://bit.ly/traceralumniunimed2021

[3] P. W. Yunanto, A. Idrus, V. M. Santi, and A. S. Hanif, “Tracer study information system for higher education,” IOP Conference Series:

Materials Science and Engineering, vol. 1098, no. 5, p. 052107, Mar. 2021, doi: 10.1088/1757-899x/1098/5/052107.

[4] Shelly Andari, Aditya Chandra Setiawan, Windasari, and Ainur Rifqi, "Educational Management Graduates: A Tracer Study from Universitas Negeri Surabaya, Indonesia," IJORER : International Journal of Recent Educational Research, vol. 2, no. 6, pp. 671–681, Nov. 2021, doi: 10.46245/ijorer.v2i6.169.

[5] A. Aminuddin, "Android Assets Protection Using RSA and AES Cryptography to Prevent App Piracy," 2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020, pp. 461–465, Nov. 2020, doi: 10.1109/ICOIACT50329.2020.9331988.

[6] Y. Nugraheni, S. Susilawati, S. Sudrajat, and A. Apriliandi, "Tracer Study Analysis of Vocational Education in Politeknik Negeri Bandung With Exit Cohort as an Approach," 2018.

[7] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[8] F. Ernawan, A. Aminuddin, D. Nincarean, M. F. A. Razak, and A. Firdaus, "Three Layer Authentications with a Spiral Block Mapping to Prove Authenticity in Medical Images," International Journal of Advanced Computer Science and Applications, vol. 13, no. 4, 2022, doi: 10.14569/IJACSA.2022.0130425.

[9] D. C. Casuat and D. E. Festijo, "Predicting Student's Employability using Machine Learning Approach," in IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) , 2019, vol. 6th.

[10] A. Binti, A. Rahman, L. Tan, and C. K. Lim, "Supervised and Unsupervised Learning in Data Mining for Employment Prediction of Fresh Graduate Students," Journal of Telecommunication, Electronic and Computer Engineering, vol. 9, pp. 2–12, 2017.

[11] B. Jonathan, P. H. Putra, and Y. Ruldeviyani, "Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE, Tomek, and SMOTE-Tomek," 2020.

[12] J. Wang, "Prediction of postoperative recovery in patients with acoustic neuroma using machine learning and SMOTE-ENN techniques," Mathematical Biosciences and Engineering, vol. 19, no. 10, pp. 10407–10423, 2022, doi: 10.3934/mbe.2022487.

[13] A. U. Umar, "Student Academic Performance Prediction using Artificial Neural Networks," International Journal of Computer Applications, vol. 178, pp. 24–29, 2019.

[14] B. Heriyadi, U. Verawardina, and T. E. Panggabean, "Tracer Study Analysis for the Reconstruction of the Mining Vocational Curriculum in the Era of Industrial Revolution 4.0 Student at Doctoral Program (S3) Vocational Education, Faculty of," 2021.

[15] A. Fernández, S. García, F. Herrera, and N. v Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," 2018.

[16] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Assessing feature selection method performance with class imbalance data," Machine Learning with Applications, vol. 6, p. 100170, Dec. 2021, doi: 10.1016/J.MLWA.2021.100170.

[17] M. Khushi et al., "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," IEEE Access, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

[18] M. Tripathi, "Sentiment Analysis of Nepali COVID19 Tweets Using NB, SVM AND LSTM," Journal of Artificial Intelligence and Capsule Networks, vol. 3, no. 3, pp. 151–168, Jul. 2021, doi: 10.36548/jaicn.2021.3.001.

[19] M. Rahardi, A. Aminuddin, F. F. Abdulloh, and R. A. Nugroho, "Sentiment Analysis of Covid-19 Vaccination using Support Vector Machine in Indonesia," IJACSA) International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, 2022, [Online]. Available: https://t.co/h5x41UO3tF

[20] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: a review," Artificial Intelligence Review, vol. 52, no. 2. Springer Netherlands, pp. 803–855, Aug. 15, 2019. doi: 10.1007/s10462-018-9614-6.

[21] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, "A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting," Annals of Data Science 2021, pp. 1–26, Jun. 2021, doi: 10.1007/S40745-021-00344-X.

[22] A. Yaqin, M. Rahardi, and F. F. Abdulloh, "Accuracy Enhancement of Prediction Method using SMOTE for Early Prediction Student's Graduation in XYZ University," International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, p. 2022, 2022, doi: 10.14569/IJACSA.2022.0130652.

[23] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," Journal of Biomedical Informatics, vol. 107, Jul. 2020, doi: 10.1016/j.jbi.2020.103465.

[24] Y. Yan, R. Liu, Z. Ding, X. Du, J. Chen, and Y. Zhang, "A parameter-free cleaning method for SMOTE in imbalanced classification," IEEE Access, vol. 7, pp. 23537–23548, 2019, doi: 10.1109/ACCESS.2019.2899467.

[25] S. Feng, J. Keung, X. Yu, Y. Xiao, and M. Zhang, "Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction," Information and Software Technology, vol. 139, p. 106662, Nov. 2021, doi: 10.1016/J.INFSOF.2021.106662.

[26] A. Aminuddin and F. Ernawan, "AuSR2: Image watermarking technique for authentication and self-recovery with image texture preservation," Computers and Electrical Engineering, vol. 102, p. 108207, Sep. 2022, doi: 10.1016/J.COMPELECENG.2022.108207.

Determining the Best Email and Human Behavior Features on Phishing Email Classification

Ahmad Fadhil Naswir[1], Lailatul Qadri Zakaria[2]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

Saidah Saad[3]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

***Abstract*—There are many email filters that have been developed for classifying spam and phishing email. However, there is still a lack of phishing email filters developed because of the complexity of feature extraction and selection of the data. There are several categories of features for classifying phishing emails, either on the email part or on the human part. The absence of which features are best for helping to classify phishing emails is one of the challenges; in the previous experiment, there was no benchmark for the features to be used for phishing email classification. This research will provide new insight into the feature selection process in the phishing email classification area. Therefore, this work extracts the features based on the category and determines which features have the most impact on classifying email as phishing or not phishing using a machine learning approach. Feature selection is one of the essential parts of getting a good classification result. Therefore, obtaining the best features from email and human behavior will significantly impact phishing classification. This research collects the public phishing email dataset, extracts the features based on category using Python, and determines the feature importance using machine learning approaches with the PyCaret library. The dataset experimented on three different experiments in which each feature category was separated, and one experiment was the combined feature selection. Binary classification is also done with the extracted features. The experiment verified that the proposed method gave a good result in feature importance and the binary classification using selected features in terms of accuracy compared to previous research. The highest result obtained is the classification with combined features with 98% accuracy. The results obtained are better compared to previous studies. Hence, this research proves that the selected features will increase the performance of the classification.**

***Keywords*—Phishing; phishing email classification; features selection; binary classification; email features; human features**

I. Introduction

In spite of the fact that numerous email filters have been created for spam emails, exceptionally few phishing email filters have been created [1]. Due to the complexity of current phishing attacks, detecting and classifying phishing attacks is a major challenge. Obtaining high-quality training data is one of the biggest problems with machine learning, as labelling data can be tedious and costly [2]. Valuating the dataset is hard because it involves figuring out the limits of the phishing email dataset and whether or not it is the same appropriate dataset as in the previous study. This is done by looking at the dataset that the previous researcher used.

For techniques used in the classification process, machine learning algorithms such as Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, and others are implemented according to the features used. From previous studies in the classification area, SVM and NB are the most commonly used methods in the phishing email classification area. The accuracy of the result from both algorithms is very satisfying. However, in this case, extraction and selection of features based on the email structures play an essential role in improving the result of email classification on specific content [3]. To improve the classification performance, a feature selection algorithm is presented, and feature selection methods are commonly used to reduce the dimensionality of datasets to improve the classification performance, reduce the processing time, or both [4]. In recent studies, there are several researches that implement word embedding for feature representation, which is one way to solve the text classification problem [5].

Phishing email classification requires clear features so that the classification produces accuracy and good performance evaluation results. The features that have been selected and extracted will represent the identity of an email itself; examples of some of the features used for email classification, especially phishing emails, are the body and URL features. Features will be extracted based on the feature type itself [6].

The relationship between features can also be determined by other fields, such as linguistic features. On the behavior side, the features extracted are classified as text feature extraction, which extracts text information that is used with the aim of representing a text message [7]. Stylometrics is one of the fields of linguistics related to the procedure for writing text, where this feature is used to identify the contents of phishing emails. Stylometric features have several categories, namely lexical, structural, content-specific, syntactic, and idiosyncratic [8]. Each category of features has its own characteristics. Email also has several main parts: a header, body, and URL.

Each corpus will be processed by following the research framework, including feature extraction. The first corpus used in this research is the IWSPA-AP 2018 dataset, which was requested by the committee of the Security and Privacy Analytics workshop [9]. The second is the custom-made corpus that combines 2 email datasets, the Enron CALO dataset and PhishCorpus [10] [11]. The detailed information about each corpus will be explained more in data collection.

Feature selection is an important stage that can affect the results of a classification process. The features to be used must have a significant impact that can make performance more accurate, especially in the machine learning area. Feature selection can be roughly classified into supervised, semi-supervised, and unsupervised methods [12]. There are standard feature selection methods for categorical data, namely Chi-Squared and information gain, but this method has drawbacks for data that has many categories. In phishing emails, some features are classified into categorical data with more than two categories. Therefore, this problem requires a new method to determine the effectiveness of the features in a dataset used. Also, there is no benchmark for the best features in the phishing email classification area [13]. PyCaret is a Python library that is useful for automating machine learning workflows. One of the features of PyCaret is feature importance, which is the process of evaluating the features that contribute the most to predicting target variables using a combination of supervised techniques, including Random Forest, Adaboost, and others [14].

The current issue regarding feature selection and extraction in the email classification area is that there is no benchmark for the feature set and which feature is the best for identifying phishing properly. Thus, it is promising that by using a combination of features on different fields with email features and using PyCaret's feature-critical algorithm can identify which features have the most significant impact on the area of classification of phishing emails. In addition, the list of best features can be produced and used as a benchmark for feature sets to help improve the performance of phishing email classification.

The rest of the paper is organized as follows: Section II discusses the works that are related to determining features. The framework for feature extraction in this experiment is in Section III, and an explanation of the data preparation and feature selection process is in Section IV. Section V will show the results of all the experiments carried out and closed with a conclusion in Section VI.

II. Related Work

This experiment is to determine the best feature for phishing email classification using feature importance, and several researchers used either stylometric features, email features, or both features for their experiments.

In [15] experiment, they proposed a classification method using the persuasion principle based on content-specific categories in the stylometric area. Several persuasion principles were used, namely: Authority, Reciprocation, and Scarcity were used as one of the feature selections in this experiment. They also used email features such as URL and body features which are included as part of feature selection. The dataset used is from Nazario PhishCorpus.

In [16], used word analysis features to detect spear-phishing emails. In contrast to the above study, this study uses a spear-phishing dataset collected from Enron Corpus because spear phishing has a specific target to attack. In the study, the analysis features used are those on the behavioral aspect, such as gender features, stylometric features, and personality features. The gender features will detect the gender of the email sender based on the choices of words in the email. The stylometric features are from the grammar side, and the personality features are emotion detection due to word selection. These features are classified as stylistic features, which is the study of the interpretation of each individual's text or spoken language in terms of accent, grammar, or word choices (lexicon). For author identification, stylistic features are often used in several journals and articles with different fields and areas, for example, author identification of a book or gender identification of a character in a novel. In this case, stylistic features are used for author identification to detect spear phishing.

Another comparison of the phishing classification using stylometric features is with [17]. The experiment extracted 26 human features more focused on syntactic feature categories. For machine learning, the classifiers used are DT, SVM, Naïve Bayes, Logistic Regression, and Neural Network. IWSPA is used as a dataset for phishing classification.

Features selection has become a crucial part of conducting email classification research. A better result will be given by selecting the best and most convenient features in the experiment. However, there are no such optimized features that can be equally applicable in all domains [18]. In the past years, the researchers tried different feature selection and extraction. The list of features to be tested is obtained based on the literature, which states that there is a lack of human features (stylometric) approaches in different research fields. In this case, stylometric features are combined with email features to detect phishing emails [19].

There are several categories from stylometrics that indicate the email is categorized as phishing or legit email. Table II shows the categories of stylometric features used in this study, namely lexical, syntactic, content-specific, structural, and idiosyncratic. The most commonly used features for phishing email classification are header, body, and URL for the email features. Features extracted are part of the main category of an email, where each category (header, body, and URL) has value in the form of text or numbers that will be analyzed at a later stage. The list of email features extracted is shown in Table I. Based on the literature survey, this research will combine features from two main features, namely human behavior focusing on stylometric features and email behavior features focusing on the structure, content, and metadata of the email itself and evaluating the effectiveness of the features extracted.

The use of PyCaret for feature engineering or classification tasks was also carried out in several experiments. The study [20] used PyCaret to focus on the feature engineering steps in the classification process using the Titanic dataset. Feature importance is used to select the best features to increase the efficiency of the classification model. PyCaret is also used for other areas, such as regression analysis. The author in [21] uses PyCaret to predict the price of a diamond. The dataset used is from the PyCaret repository. The best machine learning approaches for the experiment are Gradient Boosting Machine and Light Gradient Boosting Machine, respectively. Due to this library is newly developed in the area of machine learning

tools, there is still little research that uses PyCaret in other research fields as a supporting library.

In [22], the experiment used PyCaret to compare the selection of Polycystic Ovarian Syndrome (PCOS) attributes. The feature selection method used is GA which is the input for PyCaret. The results obtained from this experiment are accuracy of 87% with the extra tree algorithm that has been provided in PyCaret. However, the feature selection used is using another method (GA) in which Pycaret itself already has a feature for calculating feature importance which can be used to evaluate the feature that has been selected.

In [23], they proposed a classification model to detect cardiovascular disease using Pycaret. The features used have been selected in advance based on previous research in which there is no feature engineering process in this experiment. The research [24] compares 14 machine learning models from PyCaret to predict whether students will drop out or not. The results obtained by experimenting with all the features, with the Decision Tree as the most appropriate model, are pretty good. Feature importance is used to see which features affect the classification results, where feature importance is obtained according to the experimental feature analysis. Finally, experiments using PyCaret as a tool to compare models can be done well and get satisfactory results.

Clustering and classification are performed to analyze employee satisfaction using machine learning. A comparison of the best models was also carried out on 5 models included in PyCaret. The dataset used is the Kaggle-IBM analytics dataset, which consists of 1470 samples. This research uses Principal Component Analysis (PCA) for feature engineering to simplify the model features. PCA has several weaknesses, one of which is that it can eliminate information from these features because the correlation between data can be lost [25]. However, it can be seen that the flexibility of PyCaret can be applied to research in other areas as well.

It can be seen that several previous studies using PyCaret have not used the feature importance provided in PyCaret to evaluate the top features. In this study, PyCaret feature importance is used to determine the best features that can be selected to evaluate each extracted category feature.

III. Methodology

This section will determine which features significantly influence the phishing email classification process by experimenting with the dataset obtained and performing feature extraction. The results of this experiment will be a combined list of features that have a high impact on the classification of phishing emails from the email behavior part, namely the structure of email and human behavior, stylometric area. The framework for the feature extraction is shown in Fig. 1 below:

A. Data Collection

Phishing email datasets are very limited in number; there are only a few publicly available sources. In previous studies, the majority of researchers used the same dataset source, and modifications were made according to the needs of the research. In this experiment, two corpora were used to answer the question of which features had the most significant influence on the classification of phishing emails.

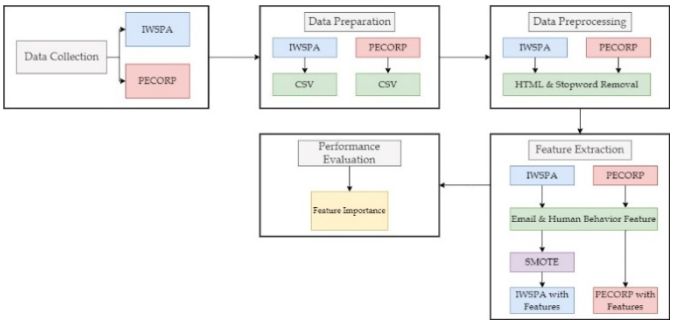


Fig. 1. The Framework for Feature Extraction

1) IWSPA-AP 2018 Corpus (IWSPA)

The first is the IWSPA-AP 2018 corpus, obtained by submitting an application to gain access to the dataset. IWSPA-AP 2018 has two different types of datasets: the IWSPA dataset with full header and no header. The full header IWSPA dataset consists of 4082 legitimate emails and 503 phishing emails, and no header IWSPA dataset consists of 5091 legitimate emails and 628 phishing emails. This corpus is classified as an unbalanced dataset because of the massive difference in the ratio of legitimate and phishing emails.

This corpus is provided in the form of text files in which every email is on a separate text file. In order to work efficiently with this data, combining all the text files into one CSV file is required to do further processing. The data extracted and transformed into CSV files is organized as follows: From, To, Date, Subject, and Body. There are additional columns, namely Label and Label Number, to determine the type of email, where 1 is for Phishing, 0 for Ham or non-phishing email.

2) Phishing Enron Corpus (PECORP)

The second corpus comes from a combination of two publicly available datasets, namely the Phishing and the Enron corpus. These two corpora are combined to create a full phishing email dataset in which the phishing emails from the Phishing corpus and legitimate emails from the Enron corpus.

A total of 2712 phishing emails come from the Online Phishing Corpus by Nazario, and 2801 legitimate emails come from the CALO Enron Email Dataset by Carnegie Mello University (CMU). This dataset was collected and prepared by the CALO Project (A Cognitive Assistant that Learns and Organizes). The CALO project dataset is the most widely used and is publicly available for download, as well as the Online Phishing Corpus by Nazario. Both corpora are combined into one CSV file. In this research, this combined corpus is called PECORP (Phishing Enron Corpus).

B. Data Preparation

Both corpora have a different format, IWSPA corpus is the text file type, and PECORP corpus is the "mbox" file type. To make the two corpora can be used as experimental material, a process is carried out to convert the two corpora into CSV format so that further processes can be carried out smoothly. Each corpus has different content of email and fields, either

from header or body that needs to be processed so that it can produce a good corpus. The corpora are converted from their original type to CSV format with the extraction of some column/fields, namely "FROM," "TO," "DATE," "SUBJECT," "BODY," and "LABEL." Both corpora's conversion and field extraction are done with Python using the PANDAS library.

The IWSPA corpus has an unbalanced amount of data ratio for the amount of data. In contrast, PECORP has been sorted for the same amount of data on phishing and legitimate emails, which was already explained in the previous section. To overcome unbalanced datasets, data processing stages are carried out so that the data in training can show good performance evaluation results. After the data preparation is complete, it will proceed to the next stage, namely data preprocessing.

C. Data Preprocessing

The preprocessing step is basically a data cleansing before the data is ready to move into the next classification process. It is crucial to preprocess the data with machine learning approaches [26]. Some preprocessing data is carried out so that the results obtained can be evaluated and meet the requirements of a good experiment. Several fields need to be preprocessed before performing feature extraction: punctuation removal for “FROM” fields, HTML checker and removal for “BODY,” and tokenization for each part of the email. The preprocessing results will be continued with the extraction of features that are in accordance with the research objective, namely human features and email features. For each corpus used, both IWSPA and PECORP will go through the preprocessing stages individually, which are carried out using the Python programming language. After this process, the data are technically feasible to pass the next stage, namely feature selection.

D. Feature Selection and Extraction

Feature selection is one stage for determining which features on the email and human side significantly affect phishing email classification. In this research, the list of features to be tested is obtained based on the literature, which states a lack of human features (stylometric) approaches in different research fields [19]. Both corpora will go through a feature extraction process after the preprocessing process has been carried out on them. The features will be extracted based on their respective categories, namely email and human behavior features. The extraction process is carried out using the Python programming language using various supporting libraries. “Pandas” library for the data frame, “re” library for regular expressions, BeautifulSoup4 for HTML file text usage, Spellchecker library for misspelt words and NLTK library for stopwords and tokenize usage. All features extracted will be placed in a new column with the appropriate data rows with the help of the Pandas library. The extracted features are as follows:

1) Email Features

Based on the observations in the literature review, the most commonly used email features for phishing email classification are header, body, and URL. Features extracted are part of the main category of an email, where each category (header, body, and URL) has value in the form of text or numbers that will be analyzed at a later stage. The process is done using Python, where each part of the email feature extraction is done in a separate function. Hence, the feature is obtained according to its category (header, body, URL).

Email feature extraction uses several Python libraries, with Jupyter Notebook as the tool. For the data frame, the Pandas library is used as the initial frame for the data analysis and manipulation. NLTK and BeautifulSoup4 libraries are used to tokenize the email field and detect HTML elements (URL and JavaScript), respectively. A regular expression is used to obtain the time from the email.

The dataset obtained and analyzed is in the form of full text, which means that the features that can be extracted are features that are in accordance with the type of data itself, for example, the number of several parts of the email, such as character length, token length, URL length, body shape, and others. Some of the URL features were extracted based on previous experiments [29]. There are several additional URL features that were extracted. The list of email features extracted is shown in Table I below:

TABLE I. LIST OF EMAIL FEATURES

Feature	Observed Field	Value	Description
Header	FROM	Char Length	Total number of characters in the “FROM” field
	SUBJECT	Token Length	Total number of tokens in the “SUBJECT” field
	TIME	Time	Time stamp when the email is received in “hour:minute” format
Body	BODY TEXT	Body Format	Boolean value that represents email body is an HTML format or non-HTML format
		JavaScript Presence	Boolean value that represents there are <script> tag in the HTML format
URL	BODY TEXT	URL Flag	Boolean value that represents the presence of URL in an email by detecting <a> tag
		URL Length	A total number of URLs character length
		URL Count	A total number of URLs found in the body text
		Presence of IP address	Boolean value that checks if the URL is on the IP address form.

2) Human Features

In terms of human features, there are 29 features from five categories that were observed and extracted. These features were extracted in each corpus, namely IWSPA and PECORP. The extracted features are based on the stylometric area, which has five categories: lexical, syntactic, structural, content-

specific, and idiosyncratic. Each category has its own characteristics which can be extracted into a feature in the body of the email. Thus, each stylometrics category produces features described descriptively in the following Table II.

TABLE II. LIST OF HUMAN FEATURES

Feature	Observed Field	Value	Description
Lexical	Full email	Character Count	A total number of characters in the email text
		Token Count	A total number of tokens in the email text
		Average Word Length	Difference between character count and token count
		Lexical Diversity	The ratio of different unique word stems (types) to the total number of words (tokens)
Syntactic	Body Text	Presence of function word	Function words include determiners, conjunctions, prepositions, pronouns, auxiliary verbs, modals, qualifiers, and question words.
Content-Specific based on [15]	Body Text	Presence of punctuation	A number of punctuations in the email text
		Authority	A total occurrence of each word of: Paypal, Verify, Fraud, Management, Identity, Debit
		Reciprocation	A total occurrence of each word of: Benefits, Bank, Customers, Accounts, Updates
		Scarcity	A total occurrence of each word of: Limited, Services, Suspension, Suspended, Terminated
Structural	Body Text	Line Count	A total number of lines in the body text
		Sentence Count	A total number of sentences in the body text
		Word Count	A total number of words in the body text
		Character Count	A total number of characters in the body text
		Average Sentence length in terms of Character	Average calculation of a total number of characters with a total number of characters
		Average Sentence length in terms of Word	Average calculation of a total number of words with a total number of words
		Average Line length in terms of Sentence	Average calculation of the total number of lines with a total number of sentences
Idiosyncratic	Body Text	Misspelt word count	Total number of possible amounts of misspelt word in the body text

The main library is the same for extracting the email features for this category, namely PANDAS and NLTK. However, several additional packages and libraries are used to extract the specific features. The feature extraction for obtaining function words is based on the syntactic feature category using the POS (Part-of-Speech) Tag method with NLTK POS Tag packages library. The packages are set to collect specific words according to function word definition (e.g., conjunction, determiners, prepositions, etc.). There is one additional package for sentence tokenization from the NLTK library. Lastly, the spellchecker python library is used for obtaining the total of a misspelt word, and the library provides the total number of possible misspelt words and the list of misspelt words.

E. SMOTE Implementation for IWSPA-2018 Corpus

The machine learning algorithm's performance is evaluated by the accuracy result and evaluation of the dataset or corpora in the experiment. The imbalanced dataset is not appropriate to get optimum results since the labelled data is not equal, and it will lead to a biased classification result [27]. There are several methods for overcoming this problem, such as random over-sampling and under-sampling, which are common approaches to solving the issue. However, these approaches have several drawbacks; under-sampling is likely to dispose of valuable data, whereas over-sampling can heighten the probability of overfitting [28]. In this research, the method used to overcome the problem regarding the imbalanced dataset is the SMOTE (Synthetic Minority Oversampling Technique) method.

SMOTE selects feature samples from the available dataset, draws a line between the samples in the feature space, and creates a new sample at a point along the drawn line. By choosing the minority class (label) for generating a new feature sample, a synthetic sample is created at a random point between the two nearest samples in the feature space [27]. IWSPA-AP 2018 has an unbalanced amount of data ratio for the amount of data, while PECORP has been sorted for the same amount of data on phishing emails and legitimate emails, which is already explained in the previous section. The balancing dataset technique is needed for the IWSPA corpus to overcome unbalanced datasets. The IWSPA-AP 2018 corpus has unbalanced data, which consists of roughly 4082 legitimate emails and 503 phishing emails. To avoid bias in the experiment result, SMOTE needs to be implemented on the IWSPA corpus, which is needed for identifying which features have the most relevant impact on phishing email classification using email and human features. For SMOTE implementation, several Python libraries and packages are required to solve the unbalanced dataset. The Imbalanced-Learn python library and SMOTE package are used to process the IWSPA corpus. By setting up the data frame that meets the requirements for the required library, the SMOTE method can be applied to the IWSPA corpus.

The balancing dataset technique is needed for the IWSPA corpus to overcome unbalanced datasets. The SMOTE technique is applied to the IWSPA corpus. The number of rows has increased from 4082 rows to 8164 rows, which means SMOTE has created feature values between each feature in the feature space. In this research, the IWSPA SMOTE will be called the IWSPA-SM corpus. Thus, the new dataset (IWSPA-SM) has been acquired and will be helpful to help determine the best feature from the selected feature set

based on email features and human features. Starting from this section down, the new dataset will be called IWSPA-SM for IWSPA SMOTE, and the old dataset will be called IWSPA-NS for IWSPA NON-SMOTE.

F. Method Implementation

The implementation of feature extraction is done using the Python programming language and is supported by Jupyter Notebook for the interface tool. The feature selection phase will determine which human and email features significantly influence the phishing email. Each corpus will undergo three experiments with different feature category selections: 1) email features, 2) human features, and 3) combining email and human features. By dividing the corpus with its extracted category features, the analysis will be able to identify which category contributed the most to the classification result. The list of experiments for feature selection is as follows:

a) IWSPA-NS (NON-SMOTE) with email features
b) IWSPA-NS (NON-SMOTE) with human features
c) IWSPA-NS (NON-SMOTE) with combined features
d) IWSPA-SM (SMOTE) with email features
e) IWSPA-SM (SMOTE) with human features
f) IWSPA-SM (SMOTE) with combined features
g) PECORP with email features
h) PECORP with human features
i) PECORP with combined features

The dataset generated from the extraction feature will be carried out to determine the importance of features for each category. Feature importance methods use various ways to obtain and calculate the feature set to determine which feature has the most impact on the current dataset. There are several types of feature importance scores, and commonly the methods are feature importance from coefficients and feature importance from a tree-based model. One way to implement feature importance is by using PyCaret, a python library that provides machine learning models for data classification, including the feature importance method. It uses a combination of several supervised techniques, including Random Forest, Adaboost, and Linear Correlation with the permutation importance technique, to select the subset of features that are most important for the model. Working with selected features instead of all the features will reduce the risk of over-fitting, improve accuracy, and decrease training time [14]. The experiment details and results are shown in the section below.

IV. Experiment and Result

This experiment has two main outputs, namely the best features in each experiment with each corpus and the performance results from binary classifications using each corpus. The feature importance result and the classification evaluation of this experiment were measured using the performance evaluation method provided by the PyCaret library; the results of the evaluation are as follows:

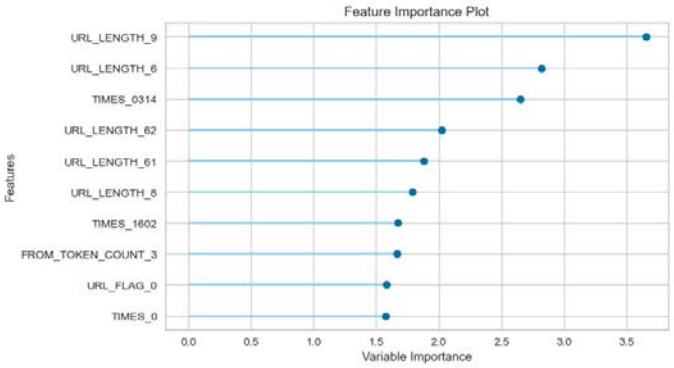


Fig. 2. IWSPA-NS Email Feature Importance Result

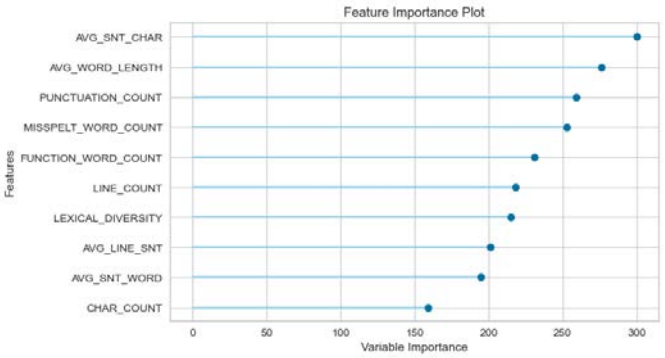


Fig. 3. IWSPA-NS Human Feature Importance Result

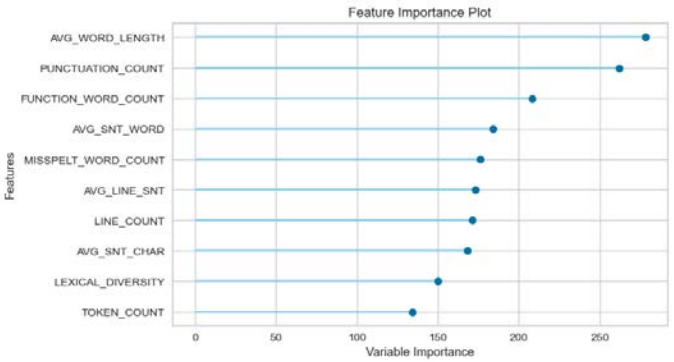


Fig. 4. IWSPA-NS Combined Feature Importance Result

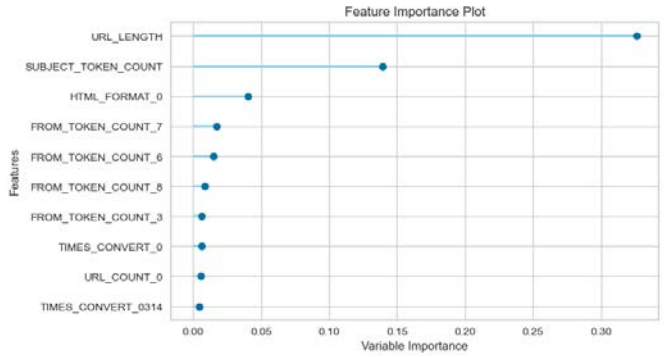


Fig. 5. IWSPA-SM Email Feature Importance Result

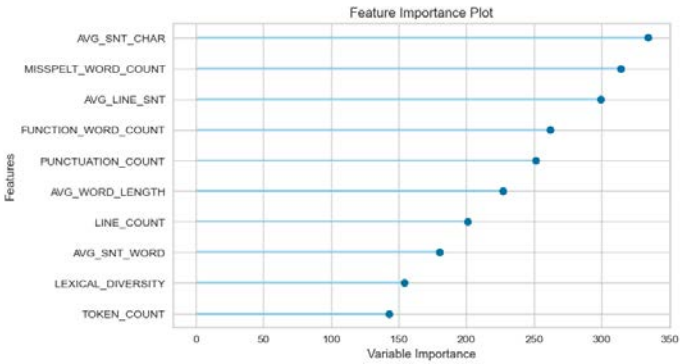


Fig. 6. IWSPA-SM Human Feature Importance Result

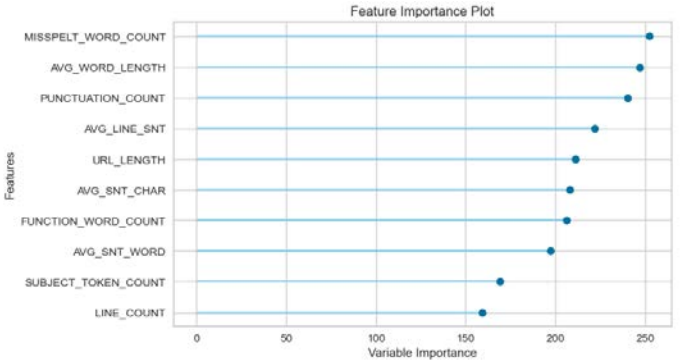


Fig. 7. IWSPA-SM Combined Feature Importance Result

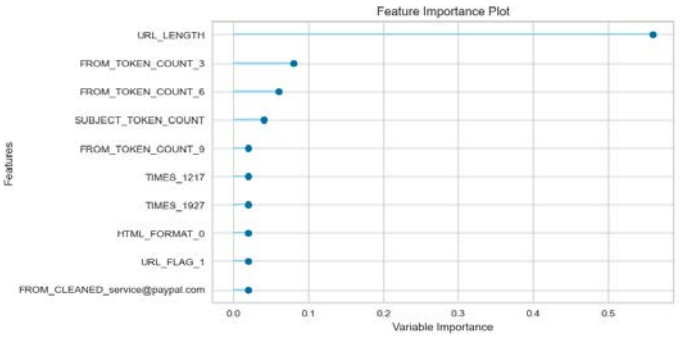


Fig. 8. PECORP Email Feature Importance Result

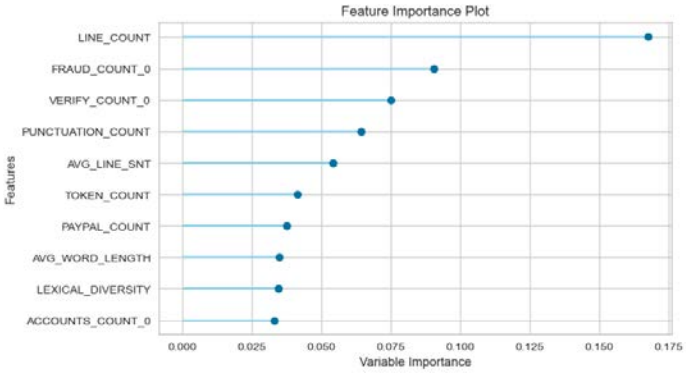


Fig. 9. PECORP Human Feature Importance Result

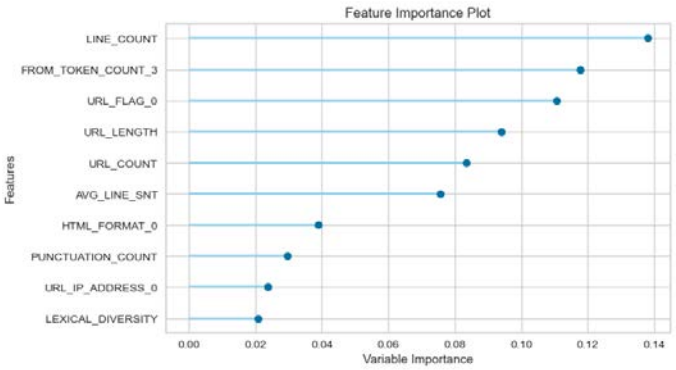


Fig. 10. PECORP Combined Feature Importance Result

The figures above are the result of the important feature of using the PyCaret library on all corpora. There are nine results obtained in each experiment where the experiment produces the output, namely the best feature that has the most significant impact in determining phishing emails in each corpus. The following are the explanations of each figure resulting from the experiment for determining the best features using feature importance with PyCaret.

Fig. 2 is the resulting diagram of IWSPA-NS with only email features extracted where the best feature is URL length. Fig. 3 is IWSPA-NS with only human features extraction, where the best feature is Average Sentence Length in terms of Characters. In Fig. 4, the best feature is the Average Word Length. In the IWSPA-SM corpus, the best features obtained are URL length for email features, Average Sentence Length in terms of characters for human features, and Misspelt Word Count for combined features, which can be seen in Fig. 5, Fig. 6, and Fig. 7 consecutively. Fig. 8 shows the result of the best feature in the PECORP corpus, namely URL length. Fig. 9 and Fig. 10 show that the best feature for experimenting with human features and combined features in the PECORP corpus is Line Count. The determination of the best feature based on the variable importance value of each feature from PyCaret uses a combination of permutation importance techniques, including Random Forest, Adaboost, and Linear correlation with the target feature. Therefore, the results obtained above are based on the algorithm of feature importance provided by PyCaret.

From the experimental results above, it can be seen that the variable importance value generated by this experiment is very diverse for several experiments on the corpus used. It can happen because the features used and extracted in each experiment are classified as "categorical features," where the coverage of the category features is extensive. For example, in the email category feature, the "Time" feature is a feature that contains numbers in time format extracted from the email header. In the human category, the "Lexical Diversity" feature contains decimal numbers with a wide range of values for each email. With a very diverse feature value of each feature extracted, the results of the variable importance value have a reasonably large range. Therefore, this experiment aims to find out what features have a significant impact on helping classify phishing emails. The scope of features that have been extracted can be in the form of numerical, boolean, or categorical values.

The results of this important feature are novelty results that can be used as a reference for the selection of features or feature engineering in the classification process using phishing email. It can be seen that some features are the same in the different corpora, for example, URL length and Line Count. This shows that the effect of these features is beneficial to improving the performance of the phishing email classification process. Moreover, further experiments can make it easier for the feature selection process to classify phishing emails with different approaches.

Tables III, Table IV, and Table V show the result of each experiment using different feature categories and corpus using the PyCaret library. The result shown above is the mean value of a 10-fold cross-validation classification with the performance metrics value for evaluation. Thirteen models are used in each classification, and the highest results from these models are shown as follows:

TABLE III. IWSPA-NS PERFORMANCE EVALUATION

Evaluation Performance	**IWSPA-NS with Email Feature**	**IWSPA-NS with Human Feature**	**IWSPA-NS with Combined Feature**
Model	Random Forest	Light Gradient Boosting Machine	Light Gradient Boosting Machine
Accuracy	0.9346	0.9698	**0.9713**
AUC	0.9001	0.9745	**0.9879**
Recall	0.4730	**0.7838**	0.7703
Precision	0.9171	0.9465	**0.9773**
F1	0.6218	0.8555	**0.8603**
Kappa	0.5904	0.8388	**0.8446**
MCC	0.6304	0.8447	**0.8528**

TABLE IV. IWSPA-SM PERFORMANCE EVALUATION

Evaluation Performance	**IWSPA-SM with Email Feature**	**IWSPA-SM with Human Feature**	**IWSPA-SM with Combined Feature**
Model	Random Forest	Light Gradient Boosting Machine	Light Gradient Boosting Machine
Accuracy	0.9107	0.9790	**0.9844**
AUC	0.9679	0.9969	**0.9982**
Recall	0.9180	0.9792	**0.9820**
Precision	0.9042	0.9786	**0.9866**
F1	0.9109	0.9789	**0.9843**
Kappa	0.8215	0.9580	**0.9688**
MCC	0.8219	0.9581	**0.9689**

TABLE V. PECORP PERFORMANCE EVALUATION

Evaluation Performance	**PECORP with Email Feature**	**PECORP with Human Feature**	**PECORP with Combined Feature**
Model	Ada Boost Classifier	Extra Trees Classifier	Decision Tree Classifier
Accuracy	0.9992	0.9964	**0.9997**
AUC	**1.0000**	0.9999	0.9997
Recall	0.9990	0.9990	**1.0000**
Precision	0.9995	0.9938	**0.9995**
F1	0.9992	0.9964	**0.9997**
Kappa	0.9984	0.9927	**0.9995**
MCC	0.9984	0.9928	**0.9995**

Table III shows the evaluation result of the experiment using the IWSPA-NS corpus with the PyCaret library. It can be seen in the comparison of the results of each category feature used in the phishing email classification process. The highest average result is in the experiment using combined features except for the recall value. Table IV shows the results of evaluating the IWSPA-SM corpus, where the highest average result was achieved in the experiment using combined features. Table V shows the experimental results of the PECORP corpus, which shows that the highest average result was obtained in the experiment using combined features. Based on the results obtained from the experiment, which are very promising, this shows that the combination of features used can improve the performance of phishing email classification. The result can be seen in Table VI.

TABLE VI. EXPERIMENT RESULT COMPARISON

Research	**Feature**	**Dataset**	**Accuracy**
IWSPA-NS Email Features	Email	IWSPA	0.9346
IWSPA-NS Human Features	Human	IWSPA	0.9698
IWSPA-NS Combined Features	Email + Human	IWSPA	**0.9713**
IWSPA-SM Email Features	Email	IWSPA	0.9107
IWSPA-SM Human Features	Human	IWSPA	0.9790
IWSPA-SM Combined Features	Email + Human	IWSPA	**0.9844**
PECORP Email Features	Email	PhishCorp + Enron	0.9992
PECORP Human Features	Human	PhishCorp + Enron	0.9964
PECORP Combined Features	Email + Human	PhishCorp + Enron	**0.9997**
Li (2020) [15]	Email + Human	PhishCorp	0.9960
Xiujuan (2019) [16]	Human	Enron	0.9505
Egozi (2018) [17]	Human	IWSPA	0.9700

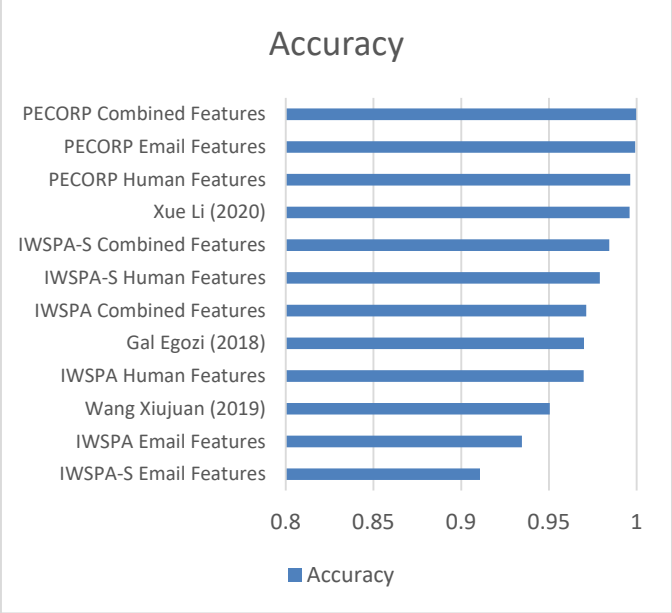


Fig. 11. Experiment Result Comparison.

A comparison of experimental results was carried out with previous related studies. In [15], this study conducted a classification of phishing emails using a combination of email and human features. The results obtained are fairly promising, but the proposed method using PyCaret is still superior, with a difference in accuracy value of 0.37%. The next comparison is with [16] and [17], where these two studies only maximize the use of human features for the classification process. With a difference in accuracy of more than 2%, the results of these studies are still just slightly below the proposed method that uses PyCaret.

Based on the comparison in Fig. 11, the experiment shows that the features selected are working best with high results even though the dataset is partly different. The Enron corpus is classified as a complex dataset because it has over 600.000+ emails on different topics and subjects. The Online Phishing corpus is more likely to ease up on preprocessing step for data analysis. Therefore, the overall comparison is categorized as a good result for this features selection experiment, especially with the combined features extraction with slightly higher accuracy than the previous research on phishing email classification with various corpora, namely [15], [16], and [17].

V. Conclusion and Future Work

By knowing which features have the most significant impact by using human and email feature extraction and selection experiments with the PyCaret library, looking at the value of the important feature for each category and corpus, and the overall high classification accuracy. Then the results of this feature selection experiment can be continued by developing embedding features that can be input for phishing email classification using a deep learning approach.

With the results of this experimental feature selection, further research can be continued using a deep learning approach for phishing email classification. The feature selection result with a high impact value on determining the phishing email classification is selected and processed for the deep learning approach by embedding the features. The feature embedding is created based on the highest feature selection, which becomes the document representation for the deep learning input. By analyzing these results, we can make a list of the features that will be used for the next step. Table VII shows the best features from the feature selection and importance experiment.

TABLE VII. Best Features

Dataset	Feature #1	Feature #2	Feature #3
IWSPA-NS Email	URL Length	Times	From Token Count
IWSPA-NS Human	Avg. Sentence by Char	Avg. Word Length	Punctuation Count
IWSPA-NS Combined	Avg. Word Length	Punctuation Count	Function Word Count
IWSPA-SM Email	URL Length	Subject Token Count	HTML Format
IWSPA-SM Human	Avg. Sentence by Char	Misspelt Word Count	Avg. Line by Sent
IWSPA-SM Combined	Misspelt Word Count	Avg. Word Length	Punctuation Count
PECORP Email	URL Length	From Token Count	Subject Token Count
PECORP Human	Line Count	“Fraud” Word Count	“Verify” Word Count
PECORP Combined	Line Count	FROM Token Count	URL Length

In Table VII above, the same features obtained from different experiments and corpora have a high impact on determining the phishing email: URL Length, Average Word Length, Average Sentence by Character, Misspelt Word, and Line Count. As a result, these features are the best features of human and email behavior for classifying phishing emails using machine learning. This feature set can become the set for experiments with phishing email classification using other approaches or as a benchmark to determine other features from human or email categories on phishing email classification using either a different dataset or the same as in this experiment.

For the next step in this research, those top selected features can be formed into a feature embedding for improving the phishing email classification results using deep learning approaches. Developing a feature representation based on the top features of each corpus and training with deep learning structures is expected to produce a better result in identifying phishing emails.

Acknowledgment

This research was supported by Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia.

References

[1] Bagui, S., D. Nandi, & S. Bagui. 2019. Classifying Phishing Email Using Machine Learning and Deep Learning. Conference: 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). 2019.

[2] Sumathi, K. & Sujatha V., 2019. Deep Learning Based-Phishing Attack Detection. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019.

[3] Mujtaba, G., Shuib, L., Raj, R. G., Majeed, N. & Al-Garadi, M. A. 2017. Email Classification Research Trends: Review and Open Issues. IEEE Access.

[4] A. Adel, N. Omar, M. Albared, & A. Al-Shabi, "Feature selection method based on statistics of compound words for Arabic text classification," Int. Arab J. Inf. Technol., 2019.

[5] S. Tiun, U. A. Mokhtar, S. H. Bakar, & S. Saad, "Classification of functional and non-functional requirement in software requirement using Word2vec and fast Text," 2020, doi: 10.1088/1742-6596/1529/4/042077.

[6] Zeeshan Bin Siddique, Mudassar Ali Khan, Ikram Ud Din, Ahmad Almogren, Irfan Mohiuddin, Shah Nazir, "Machine Learning-Based Detection of Spam Emails", Scientific Programming, vol. 2021, Article ID 6508784, 11 pages, 2021. https://doi.org/10.1155/2021/6508784

[7] M. Suhaidi, R. Abdul Kadir, & S. Tiun, "A REVIEW OF FEATURE EXTRACTION METHODS ON MACHINE LEARNING," J. Inf. Syst. Technol. Manag., 2021, doi: 10.35631/jistm.622005.

[8] N. M. Sharon Belvisi, N. Muhammad, & F. Alonso-Fernandez, "Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features," 2020, doi: 10.1109/IWBF49977.2020.9107953.

[9] R. M. Verma, V. Zeng, & H. Faridi, "Poster: Data quality for security challenges: Case studies of phishing, malware and intrusion detection datasets," 2019, doi: 10.1145/3319535.3363267.

[10] Electronic Discovery Reference Model. Enron Email Corpus CALO version. 2010. Available: https://www.cs.cmu.edu/~./enron/

[11] Nazario, J., 2006. Online Phishing Corpus. Available: https://monkey.org/~jose/phishing/

[12] Miao, J. & Niu, L. "A Survey on Feature Selection," 2016, doi: 10.1016/j.procs.2016.07.111.

[13] Brownlee, J. "Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python,". Machine Learning Mastery, 2020.

[14] Ali M., "PyCaret," PyCaret: An open source, low-code machine learning library in Python, 2020. Available: https://www.pycaret.org

[15] Li, X., D. Zhang, & B. Wu, "Detection method of phishing email based on persuasion principle," 2020, doi: 10.1109/ITNEC48623.2020.9084766

[16] Xiujuan, W., Chenxi, Z., Kangfeng, Z., & Haoyang, T., 2019. Detecting Spear-phishing Emails Based on Authentication. IEEE 4th International Conference on Computer and Communication Systems. 2019.

[17] Egozi, G. & Verma R., "Phishing email detection using robust NLP techniques," 2019, doi: 10.1109/ICDMW.2018.00009.

[18] Iqbal, F., Khan, L. A., Fung, B. C. M. & Debbabi, M. 2010. E-mail authorship verification for forensic investigation. Proceedings of the ACM Symposium on Applied Computing.

[19] K. Lagutina et al., "A Survey on Stylometric Text Features," 2019, doi: 10.23919/FRUCT48121.2019.8981504.

[20] An, S. "How to use PyCaret with Feature Engineering". Accessed 2021 from https://www.kaggle.com/code/subinium/how-to-use-pycaret-with-feature-engineering/

[21] Ali M. "Introduction to Regression in Python with PyCaret". Accessed December 12, 2021 from https://towardsdatascience.com/introduction-to-regression-in-python-with-pycaret-d6150b540fc4

[22] Munjal, A., Khandia, R., and Gautam, B., "A Machine Learning Approach for Selection of Polycystic Ovarian Syndrome (PCOS) Attributes and Comparing Different Classifier Performance with the help of Weka and Pycaret," Int. J. Sci. Res., 2020, doi: 10.36106/ijsr/5416514.

[23] Urmila, P. & Tejashree, S. 2021. Automl: Building An Classfication Model With Pycaret. Ymer. 20. 547-552.

[24] Anwar, M. T, and Permana, D. R. A., "Perbandingan Performa Model Data Mining untuk Prediksi Dropout Mahasiwa," J. Teknol. dan Manaj., 2021, doi: 10.52330/jtm.v19i2.34.

[25] I Ketut Adi, W. & Handri, S. (2022). Analisis Employee Satisfaction Menggunakan Teknik Clustering Dan Classification Machine Learning. Jurnal Ilmiah Informatika Komputer. 18. 10. 10.35889/progresif.v18i1.766.

[26] Ali M. H & Lailatul Q. Z. "Question Classification Using Support Vector Machine and Pattern Matching," Journal of Theoretical and Applied Information Technology, 20th May 2016. Vol.87. No.2. 2016.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., 2002, doi: 10.1613/jair.953.

[28] A. Y. Taha, S. Tiun, A. H. A. Rahman, and A. Sabah, "Multilabel Over-sampling and Under-sampling with Class Alignment for Imbalanced Multilabel Text Classification," J. Inf. Commun. Technol., 2021, doi: 10.32890/JICT2021.20.3.6.

[29] Ahmad F. N., Lailatul Q. Z, Saidah S., "The Effectiveness Of URL Feature on Phishing Emails Classification using Machine Learning Approach". Asia-Pacific J. Inf. Technol. Multimed., June 2022.

A Scalable Machine Learning-based Ensemble Approach to Enhance the Prediction Accuracy for Identifying Students at-Risk

Swati Verma[1], Rakesh Kumar Yadav[2], Kuldeep Kholiya[3]
Research Scholor, IFTM University Moradabad, Uttar Pradesh, India[1]
Assistant Professor, IFTM University Moradabad, Uttar Pradesh, India[2]
Assistant Professor, B.T. Kumaon Institute of Technology, Dwaraht, Uttarakhand, India[3]

***Abstract*—Among the educational data mining problems, the early prediction of the students' academic performance is the most important task, so that timely and requisite support may be provided to the needy students. Machine learning techniques may be used as an important tool for predicting low-performers in educational institutions. In the present paper, five single-supervised machine learning techniques have been used, including Decision Tree, Naïve Bayes, k-Nearest-Neighbor, Support Vector Machine, and Logistic Regression. To analyze the effect of an imbalanced dataset, the performance of these algorithms has been checked with and without various resampling methods such as Synthetic Minority Oversampling Technique (SMOTE), Borderline SMOTE, SVM-SMOTE, and Adaptive Synthetic (ADASYN). The Random hold-out method and GridSearchCV were used as model validation techniques and hyper-parameter tuning respectively. The results of the present study indicated that Logistic Regression is the best performing classifier with every balanced dataset generated using all of the four resampling techniques and also achieved the highest accuracy of 94.54% with SMOTE. Furthermore, to improve the prediction results and to make the model scalable, the most suitable classifier was integrated with the help of bagging, and a well-accepted accuracy of 95.45% was achieved.**

***Keywords*—Educational data mining; resampling methods; feature selection technique; machine learning; imbalanced data**

I. Introduction

Due to the digitization and use of technology in the educational field, there is a large amount of educational data. Educational Data Mining helps to analyze and extract useful information, such as selecting the factors that affect the students' performance, predicting students' performance, etc., from a large amount of educational data. As students or youths are the future of any nation, predicting the success rate of students in their academic area is a very important and beneficial task. This may be achieved with the help of educational data mining, which utilizes various machine learning techniques.

Although the field of Educational Data Mining (EDM) is old and its definition was given by Fayyad et al. [1] in 1996, EDM emerged as a convincing research area after the establishment of the annual International Conference on Educational Data Mining and the Journal of Educational Data Mining in 2008 [2]. After that, Baker [3] identified the application of data mining in education to discover models for predicting students' performance by using the methods of prediction, clustering, relationship mining, and discovery with models. Among the applications of EDM, detecting student failure at an early stage has been an appealing research topic for researchers due to its social impact. The prediction of the students at risk of being dropouts from an institute or school becomes difficult due to the large number of factors that may influence the academic performance of the students. Thus, it is quite important to predict low-performing students at an early stage with higher accuracy, along with the important factors that may affect their performance.

To achieve this goal, the present study has three important research objectives: (i) to identify the influential features by using a filter-based feature selection technique. (ii) to identify the best performing classifier by comparing various single-supervised machine learning techniques, viz., decision trees, Naïve Bayes, k-Nearest Neighbor, Logistic Regression, and Support Vector Machine with various resampling techniques such as random oversampling, SMOTE, Borderline SMOTE, SVM-SMOTE, and ADASYN. (iii) to enhance the prediction rate of the students at-risk by using an ensemble model that integrates the most suitable data mining technique.

In rest of the paper, the work related to the present study is given in section II. The methodology used in the present work is explained in Section III. In Section IV, the obtained results are analyzed and discussed. Finally, the conclusion and future work are given in Section V.

II. Related Work

In the past, various review studies have been performed on educational data mining [4, 5], and many researchers have worked on identifying the factors that deteriorate the academic performance of students. Ahmed et al. [6] selected nine attributes such as department, attendance, high school degree, mid-term marks, student participation, lab test grades, assignment scores, seminar performance, and homework to predict the final grade and generate the rules set by the Decision Tree. Tomasevic et al. [7] have compared the performance of several data mining algorithms using past student performance, student engagement, and student demographic data. They concluded that students' engagement and past performance data have a significant influence, while

demographic attributes have a slight impact, on students' performance. Further, Verma and Yadav [8] used the cross-tabulation method and the chi-square test to analyze the effects of different attributes such as background, academic, social, and psychological characteristics on students' academic performance. In their finding, it was concluded that students' academic and background attributes were the most influential factors that may affect students' grades.

With knowledge of the factors that influence the students' performance, predictions can be made with the help of data mining algorithms to identify students at risk. To analyze students' performance, Asif et al. [9] implemented decision tree and clustering technique on a dataset of 210 students that contained pre-admission marks and all subjects' marks and found that the pre-university marks and subjects' marks in the first and second years had an impact on students' final year marks. Hamoud et al. [10] applied Bayesian classifiers, namely Naïve Bayes and Bayes Net, to the dataset of 161 students and found that Naïve Bayes outperformed for predicting the students' performance. Costa et al. [11] performed a comparison of the effectiveness of different educational data mining techniques to predict students' performance in introductory programming courses and concluded that the support vector machine outperformed. Moreover, Ha et al. [12] implemented rule-based learners, neural-based learners, and statistical-based learners (Naïve Bayes, and Support Vector Machine) on students' datasets, which consist of personal and past academic information, to predict students' performance. In their experiment, neural-based learners and Naïve Bayes achieved the highest accuracy of 86.19%.

A suitable approach towards feature selection and handling imbalanced class problems may enhance the prediction accuracy of machine learning models. Thammasiri et al. [13] compared random oversampling and SMOTE balancing methods along with four popular data mining models: logistic regression, decision trees, neural network, and support vector machine to assess the students' performance. In their results, Support Vector Machine (SVM) achieved the highest accuracy of 90.24% with SMOTE. Mueen et al. [14] applied Naïve Bayes, Neural Network, and Decision Tree to students' data having their general, academic, and forum-related variables along with feature selection and SMOTE oversampling method to solve the imbalanced data problem and found Naïve Bayes to be outperformed with 86% accuracy. Ghorbani and Ghousi [15] used and compared different resampling methods, viz., Borderline SMOTE, Random Over Sampler, SMOTE, SMOTE-ENN, SVM-SMOTE, and SMOTE-Tomek, by evaluating the performance of the various classifiers, and Random Forest obtained the highest accuracy of 81.27% with SVM-SMOTE. Further, Ghavidel et al. [16] solved the problem of imbalanced data by using a combination of the SVM-SMOTE (an over-sampling technique) and Edited-Nearest-Neighbor (an under-sampling technique) while predicting disease mortality. Recently, Desiani et al [17] applied k-Nearest Neighbor (k-NN), Artificial Neural Network (ANN), and C4.5 to students' educational background records along with SMOTE to make the dataset balanced, and that balanced dataset increased the accuracy of prediction, and for k-NN the maximum achieved accuracy was 83.71%.

Another aspect that enhances the prediction accuracy is the appropriate use of ensemble models. Teoh et al. [18] used feature selection and SMOTE oversampling techniques and then applied various ensemble machine learning methods, namely stacking, boosting, and bagging. In their findings, AdaBoost has achieved a maximum accuracy of more than 90%.

Although there are several studies to predict the students' academic performance, the study which considers all categories of variables, i.e., background, academic, social, and psychological, and predicts students at-risk at an early stage with adequate accuracy is lacking. Also, a single classifier-based prediction is not suitable from one perspective to another. Moreover, a classifier giving the highest prediction accuracy for a particular dataset may not be valid for a different dataset. Thus, the aim of the present study is to identify low performers at an early stage with a higher prediction rate by using a scalable approach.

III. METHODOLOGY

The main objective of the present paper is to predict the academic performance of students with higher accuracy. To achieve this goal, the different single supervised machine learning algorithms were applied with and without data balancing, and finally, by comparing the results, a model was constructed to enhance the prediction accuracy. The methodology applied in the present work may be given as follows:

- Dataset preparation.
- Data preprocessing including data transformation, feature selection, and data balancing.
- Identification of the best classification technique by comparing the results of classification models when applied to the preprocessed data.
- Make a scalable ensemble model with the help of the best classification technique.
- Result evaluation of the proposed ensemble model.

The workflow of the proposed methodology is given in Fig. 1.

A. Dataset

To make the data versatile, it is collected from the two different engineering colleges situated in different regions (the north and south of India). In the present paper, the sample size comprises 550 engineering students from two different engineering colleges in India, i.e., Bipin Tripathi Kumaon Institute of Technology, Dwarahat, Uttarakhand, and Cochin University of Science & Technology, Trivandrum, Kerala. The dataset includes information regarding background, past academic, social, and psychological factors with 30 different attributes, of which three attributes (roll-number, name, and branch) are used for identification purposes only and do not play any role in the prediction of low-performers. So, only 27 attributes were used for the present work, with first semester GPA as the output variable. For these attributes, data was collected online with the help of a multiple-

choice questionnaire created via outsourced technology, i.e., Google Form. As the aim of the paper is to identify the students having the highest risk of dropping out of college, the information about the output attribute for the dataset is divided only into two categories, i.e., low performers and high performers, based on the first-semester grade point of the students.

B. Data Preprocessing

Before applying any machine learning model to the dataset, data should be preprocessed so that any machine learning model can be performed efficiently. In the present study, the dataset is complete and free from noise, so there is no need to handle missing data and outliers. To preprocess the data, data transformation, feature selection, and data balancing have been performed.

1) Data transformation: In the present study, all the features were categorical except students' GPA as it was initially in numerical form. So, GPA was generalized into categorical values, i.e., "class A (high performer)" and "class B (low performer)". Finally, these categorical variables were encoded into the suitable format of machine learning models.

2) Feature selection: Feature selection is an important part of the students' performance prediction model for two main reasons:

- The main purpose of the prediction of students' academic performance is to provide timely support to the low-performing students in the area where they are lacking. Only after identifying the attributes that have a significant impact on the output variable, i.e., students' academic performance, suitable corrective measures may be taken to provide support to the low-performing students.
- With the help of feature selection, irrelevant attributes may be removed from the data without losing reliability in classification. Thus, the dimensionality reduction raises the processing speed, and hence the classifier can learn faster.

There are three main feature selection techniques: manual selection based on pedagogical theories or expert experience; filter-based selection; and wrapper feature selection [19]. In the present study, as all the attributes were categorical, a filter-based feature selection technique, namely "chi-square", was used by which p-values were calculated for each attribute [8]. The attributes having a p-value of less than 0.01 show a highly significant correlation with the student's grades.

3) Data balancing: Data balancing is an important part of preprocessing step by which class distribution have to make equal so that classifier do not assign every new sample to the majority class only. In the present study the distribution of "class A" and "class B" is shown in Fig. 2. From the figure, it may be revealed that the dataset contained more samples from "class A" (66%) than the "class B" (34%). Previous study [20] shows that if the percentage of minority class is less than 35% of dataset then it is called imbalanced and hence the dataset of present study is imbalanced to some extent. There are mainly three types of re-sampling techniques i.e., over-sampling, under-sampling, and hybrid-sampling [15] that may be used to balance the dataset. Due to the limited size of dataset, in the present study, only over-sampling techniques i.e., Synthetic Minority Oversampling Technique (SMOTE) [21], Borderline SMOTE [22], SVM-SMOTE [23], and ADASYN [24] were used and compared.

C. Machnie Learning Techniques

There are different types of classification machine learning models that may be used to predict the students' academic performance. In the present study, five single supervised machine learning models have been applied, including Decision Tree [25], Naïve Bayes [9, 26], k-Nearest-Neighbor [27], Support Vector Machine [28], and Logistic Regression [29]. To achieve the best performance of these machine learning models, the passing parameters for these models were set with the help of an algorithm called "GridSearchCV" which gives the best combination of passing parameters [30]. These combinations of passing parameters are listed in Table I.

TABLE I. CLASSIFICATION MODELS AND THEIR PASSING PARAMETERS

Machine learning model	**Passing parameters**
Decision Tree	Criterion="gini", max_depth=4, max_leaf_node=8
Naïve Bayes	No parameter
k-Nearest Neighbor	n_neighbor=21
Support Vector Machine	c=2, kernel="rbf"
Logistic Regression	No parameter

D. Model Validation and Result Evaluation

Model validation is used to check the effectiveness of the model across independent datasets. In the present study, the random hold-out method was used for model validation, in which 80% of the data was for training purposes and 20% of the data was reserved for testing purposes.

Furthermore, the performance of all the machine learning techniques was evaluated in terms of accuracy, precision, recall, and f1-score. These performance metrics are given as follows:

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{Total\ number\ of\ samples} \quad (1)$$

$$Precision = \frac{True\ Positive}{Total\ classes\ predicted\ as\ positive} \quad (2)$$

$$Recall = \frac{True\ Positive}{Total\ number\ of\ actual\ positive\ classes} \quad (3)$$

$$F1 - score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

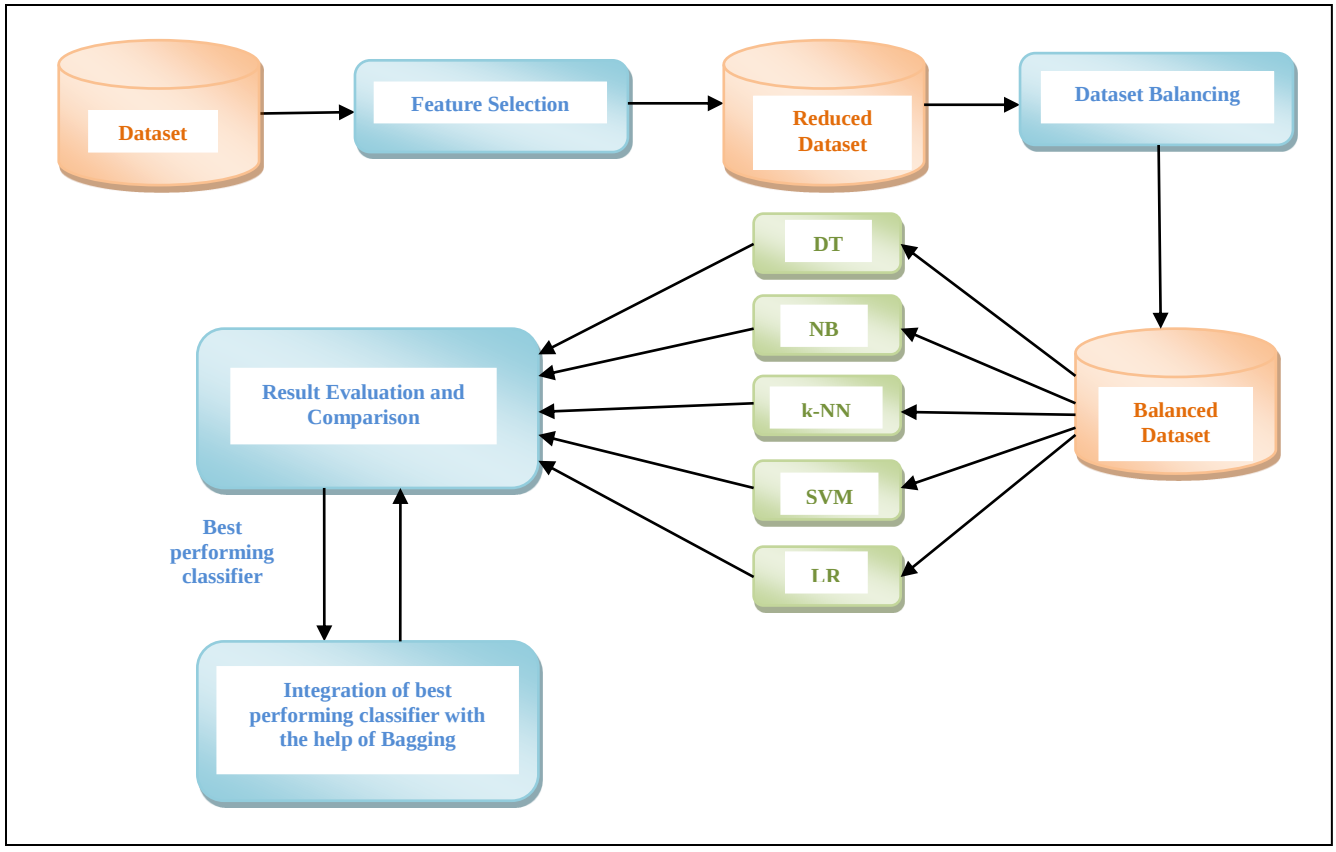


Fig. 1. Framework of Proposed Methodology.

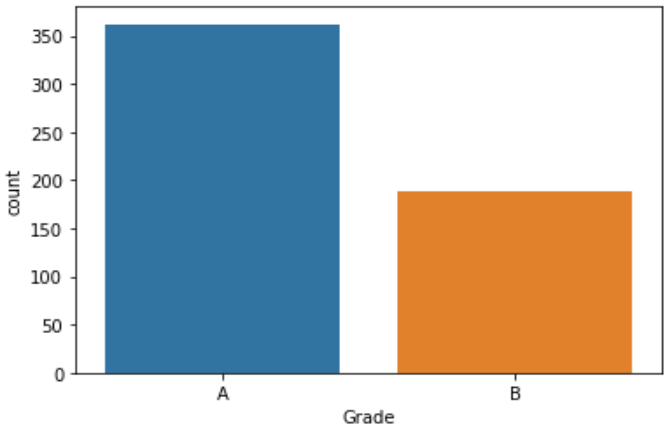


Fig. 2. Distributions of the Grades.

E. Construction of Ensemble based Classifier

In most of the previous studies [18, 31–38], it was shown that the ensemble model gives a higher prediction accuracy, so, to enhance the prediction accuracy, an ensemble model was constructed in the present study. For this, the best performing classifier was selected along with its suitable resampling method, after comparing the results of different single machine learning algorithms with balanced dataset. Finally, in order to make an ensemble classifier, the three best-performing classifiers were integrated with the help of bootstrap aggregation.

IV. Result and Discussion

In the present work, the whole experiment was done with the help of different libraries such as Pandas, Seaborn, and Scikit-learn of the Python programming language, which is a very powerful and user-friendly language for data scientists. The first aspect of the present work is to find out the influential attributes and to reduce the dimensionality with the help of a filter-based feature selection technique. For this purpose, the p-values were calculated for different attributes using the chi2 method of the sklearn.feature_selection library of Python programming and are shown in Table II. From this table, it is depicted that after applying the feature selection technique, the following 11 features are selected as influential features that affect students' academic performance: percentage in 10^{th} standard, percentage in 12^{th} standard, confidence, mathematics % in 12^{th} standard, punctuality, curiosity, medium/language of previous study, category, father's highest qualification, mother's highest qualification, and mental stress.

After selecting the most influential attributes, Decision Tree, Naïve Bayes, k-Nearest-Neighbor, Support Vector Machine, and Logistic Regression algorithms have been applied to the dataset, which contains only the 11 selected most influential attributes. The results obtained for accuracy, precision, recall, and f1-score of these algorithms are represented in Table III.

TABLE II. STUDENTS' RELATED INPUT FEATURES AND THEIR CORRESPONDING P-VALUES

Attribute Category	Attribute	p-value
Background Attributes	Gender	.0304
	Category	8.1425e-05
	Number of Siblings	.5330
	Status of Parent	.4112
	Father's Highest Qualification	.0001
	Mother's Highest Qualification	.0027
	Father's Occupation	.8812
	Mother's Occupation	.8034
	Annual Family Income	.2393
	Living Location	.1042
	Medium/Language of Previous Study	4.4161e-06
Academic Attributes	Percentage in 10^{th} standard	1.0815e-42
	Percentage in 12^{th} standard	1.3151e-35
	Entrance Exam/JEE Rank	.1319
	Average Self-Study Time	.0407
	Mathematics % in 12^{th} standard	2.9478e-29
Social Attributes	Participation in Extra-Curricular Activities	.4782
	Whether have Friends	.9547
Psychological Attributes	Motivation to Join Course	.9281
	Mental stress	.0033
	Homesickness	.2046
	Personality	.1333
	Adaptability	.4372
	Confidence	6.5301e-33
	Curiosity	2.0818e-09
	Punctuality	5.1669e-14

TABLE III. RESULTS OF THE CLASSIFIERS ON IMBALANCED DATASET

Classifier	Accuracy (in %)	Recall		Precision		F1-score	
		A	*B*	*A*	*B*	*A*	*B*
Decision Tree	91.81	0.99	0.79	0.90	0.97	0.94	0.87
Naïve Bayes	88.18	0.89	0.87	0.93	0.80	0.91	0.84
k-Nearest Neighbor	89.09	0.94	0.79	0.89	0.88	0.92	0.83
Support Vector Machine	90.90	0.99	0.76	0.89	0.97	0.93	0.85
Logistics Regression	92.72	0.99	0.82	0.91	0.97	0.95	0.89

From Table III, it may be observed that the highest accuracy, i.e., 92.72%, was achieved with Logistic Regression. In terms of recall and precision for classes A and B, no single algorithm can be declared best. This is because precision and recall for classes A and B are not the highest for the same algorithm. For example, in Naïve Bayes recall and precision for class B and class A is highest, respectively, but recall for class A and precision for class B is lowest. In such situations, the f1-score may be taken as an evaluation criterion, as the f1-score is the harmonic mean of precision and recall. Logistic Regression has achieved the highest accuracy and highest f1-score for both classes 'A' and 'B', and hence it may be considered the best performing algorithm with the imbalanced dataset. The dataset of the present study was imbalanced, and hence four resampling techniques (SMOTE, Borderline SMOTE, SVM-SMOTE, and ADASYN) have been used, and the performance of all the classifiers was evaluated with the balanced dataset.

The performances of different models with the different resampling methods are shown in Table IV. From Table IV, it may be noted that the accuracy of the models, except for Logistic Regression, was not significantly improved when applied to the balanced dataset. This may be because of the fact that, in the case of balanced data, all the algorithms considered both the classes "A" and "B" with equal weightage. So, it may be concluded that although in the case of balanced datasets, the accuracy of every classifier is not increasing; the prediction accuracy may now be trustable and sufficient to measure the model's performance. The performances of various classifiers using the resampling methods SMOTE, Borderline SMOTE, SVM-SMOT, and ADASYN are shown in Fig. 3-6 respectively. From these figures, it may be observed that Logistic Regression outperformed all the classifiers in every balanced dataset generated with all the four resampling techniques, and the highest accuracy of 94.54% and the highest F1-score were achieved when SMOTE was considered as a resampling method.

TABLE IV. RESULTS OF THE CLASSIFIERS ON BALANCED DATASET

Classifier	Evaluation Metric		SMOTE	Borderline SMOTE	SVM- SMOTE	ADASYN
Decision Tree	Accuracy (in %)		89.09	88.18	88.18	91.81
	Recall	A	0.89	0.86	0.89	0.92
		B	0.89	0.92	0.87	0.92
	Precision	A	0.94	0.95	0.93	0.96
		B	0.81	0.78	0.80	0.85
	F1-score	A	0.91	0.91	0.91	0.94
		B	0.85	0.84	0.84	0.89
Naïve Bayes	Accuracy (in %)		80.90	83.63	86.36	83.63
	Recall	A	0.86	0.83	0.85	0.83
		B	0.71	0.84	0.89	0.84
	Precision	A	0.85	0.91	0.94	0.91
		B	0.73	0.73	0.76	0.73
	F1-score	A	0.86	0.87	0.89	0.87
		B	0.72	0.78	0.82	0.78
k-Nearest Neighbor	Accuracy (in %)		85.45	82.72	83.63	81.81
	Recall	A	0.86	0.79	0.79	0.78
		B	0.84	0.89	0.92	0.89
	Precision	A	0.91	0.93	0.95	0.93
		B	0.76	0.69	0.70	0.68
	F1-score	A	0.89	0.86	0.86	0.85
		B	0.80	0.78	0.80	0.77
Support Vector Machine	Accuracy (in %)		90.90	90.00	89.09	90.90
	Recall	A	0.96	0.92	0.92	0.94
		B	0.82	0.87	0.84	0.84
	Precision	A	0.91	0.93	0.92	0.92
		B	0.91	0.85	0.84	0.89
	F1-score	A	0.93	0.92	0.92	0.93
		B	0.86	0.86	0.84	0.86
Logistics Regression	Accuracy (in %)		94.54	90.90	91.81	93.63
	Recall	A	0.99	0.93	0.94	0.97
		B	0.87	0.87	0.87	0.87
	Precision	A	0.93	0.93	0.93	0.93
		B	0.97	0.87	0.89	0.94
	F1-score	A	0.96	0.93	0.94	0.95
		B	0.92	0.87	0.88	0.90

TABLE V. RESULTS OF THE PROPOSED MODEL

Classifier	Evaluation Metric		Imbalanced dataset	SMOTE	Borderline SMOTE	SVM- SMOTE	ADASYN
Proposed Model	Accuracy (in %)		93.63	95.45	93.63	93.63	94.54
	Recall	A	0.99	0.99	0.96	0.96	0.97
		B	0.84	0.89	0.89	0.89	0.89
	Precision	A	0.92	0.95	0.95	0.95	0.95
		B	0.97	0.97	0.92	0.92	0.94
	F1-score	A	0.95	0.97	0.95	0.95	0.96
		B	0.90	0.93	0.91	0.91	0.92

Finally, after evaluating the performance of all classifiers, the best performing classifier, namely Logistic Regression, was chosen to create the ensemble model in order to improve prediction accuracy. In order to make the ensemble model, three Logistic Regression classifiers were integrated with the help of bagging. The result of the proposed integrated model is shown in Table V. The proposed model has achieved the highest accuracy of 95.45%, the highest prediction rate for low performers, and the highest f1-score for both classes while using SMOTE. It is pertinent to mention here that the accuracy of the proposed model increased by 1.82% after using the resampling technique SMOTE, while in the study of Desiani et al., the average accuracy was increased by 20.13%. The possible reason may be that the dataset used in the present study has a small sample size and was not highly imbalanced. In the case of a large sample size, the number of students at risk will be significantly lower, and hence, in such situations of highly imbalanced data, the present model may be quite useful.

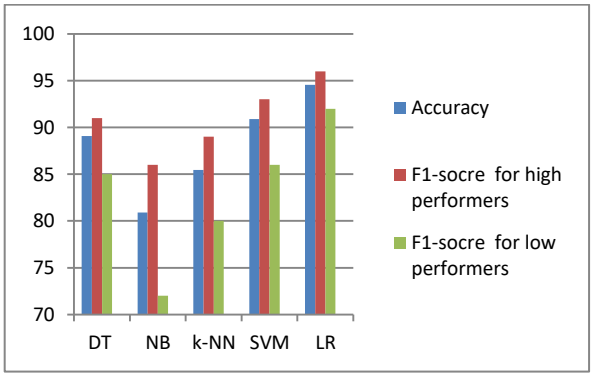


Fig. 3. Performance of Different Classifiers with SMOTE.

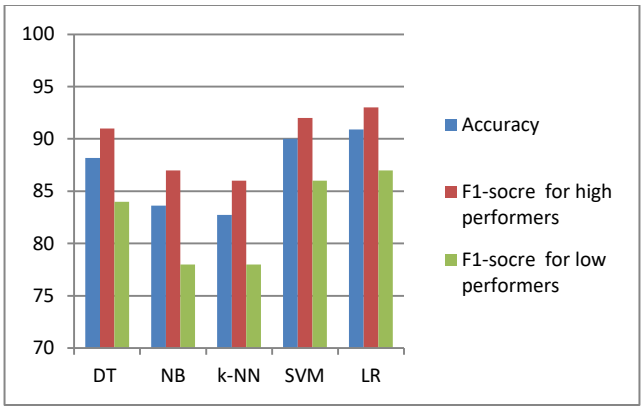


Fig. 4. Performance of Different Classifiers with Borderline SMOTE.

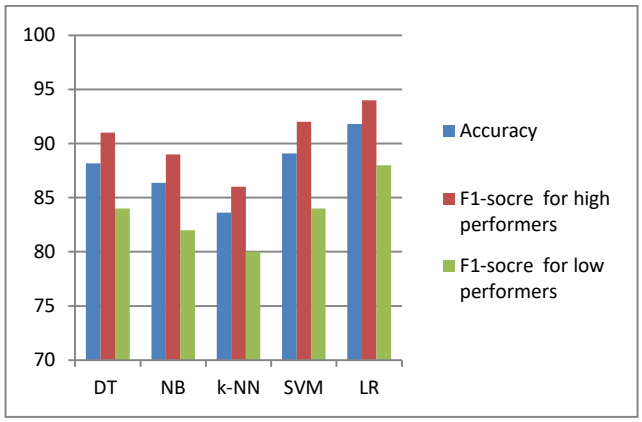


Fig. 5. Performance of Different Classifiers with SVM-SMOTE.

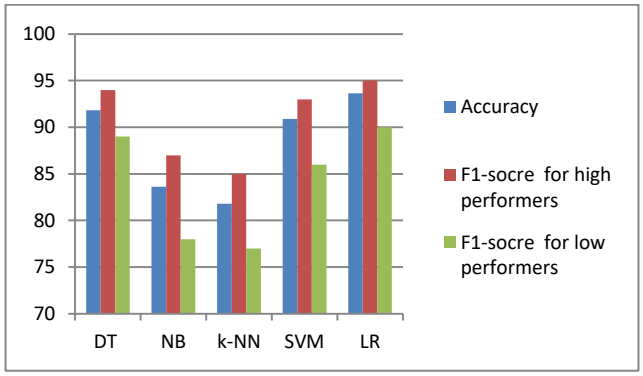


Fig. 6. Performance of Different Classifiers with ADASYN.

The highest prediction accuracy achieved in the present study is 95.45%, which is greater than most of the previous studies [12-18]. Along with the enhanced prediction accuracy, the main advantage of the present work is that the methodology proposed in the present study is scalable from one context to the other.

V. Conclusion and Future Work

From the present work, it may be concluded that students' past academic performance (10th standard %, 12th standard %, and Math's % in the 12th standard), their background (category, parents' qualification, and medium of the previous study), and their psychological features (mental stress, confidence, curiosity, and punctuality) were the relevant attributes. Thus, to increase the academic performance of the students, these factors may be considered as the focus points.

In the present study, all the used classifiers were able to predict students' outcomes with reasonable accuracy of more than 80%. Among all the used classifiers, Logistic Regression was the best performing algorithm with a balanced as well as an imbalanced dataset. Further, the accuracy and prediction rate for identifying low performers as well as for high performers were improved when the Logistic Regression was applied to the balanced dataset. The prediction accuracy was further enhanced with the use of an ensemble classifier in which three Logistic Regression classifiers (because of its highest performance) were integrated with the help of bootstrap aggregation. The proposed integrated model has achieved the highest accuracy of 95.45% and the highest precision and recall for low performers with the balanced dataset formulated with the help of the resampling technique SMOTE.

It should be noted that with different datasets, the different classifiers may give the highest prediction accuracy, and hence there is a need for the methodology to be scalable for every situation. Thus, the main advantage of the present approach is its scalability for different datasets. Further, this study may also be applied to the different domains of data mining and machine learning applications for enhancing prediction accuracy. The limitation of the present study is that the examined dataset has a small sample size and slightly imbalanced data, so in the future, the proposed methodology should be used with large sample sizes and highly imbalanced data for the prediction of students' academic performance.

References

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI magazine, Vol. 17, pp. 37-53, 1996.

[2] R. S. Baker, "Educational Data Mining: An Advance for Intelligent Systems in Education", IEEE Intelligent systems, 29, 78–82, 2010.

[3] R. S. J. D. Baker. "Data Mining for Education", International Encyclopaedia of Education, 3rd edition, Vol. 7, pp. 112-118, 2010.

[4] B. Albreiki, N. Zaki, and H. Alashwal, "A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques", Educational Sciences, Vol. 11, No. 9, pp. 1-27, 2021.

[5] W. Xiao, P. Ji, and J. Hu, "A survey on educational data mining methods used for predicting students' performance", Engineering Reports, pp. 1-23, 2021.

[6] A. Ahmed, and I. Elaraby, "Data mining: A prediction for student's performance using classification method", World Journal of Computer Application and Technology, Vol. 2, pp. 43-47, 2014.

[7] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," Computers & Education, Vol. 143, pp. 1-18, 2020.

[8] S. Verma, and R. K. Yadav, "Effect of Different Attributes on the Academic Performance of Engineering Students", ICATMRI, pp. 1-4, 2020.

[9] R. Asif, A. Merceron, S. A. Ali,and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Computers & Education, Vol. 113, pp. 177-194, 2017.

[10] A. K. Hamoud, A. M. Humadi, W. A. Awadh, and A. S. Hashim, "Students' Success Prediction based on Bayes Algorithm," International Journal of Computer Application, Vol. 178, No. 7, pp. 6-12, 2017.

[11] E. B. Costa, B. Fonseca, M. A. Santana, F. Araújo, and J. Rego,

"Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming course", Computers in Human Behavior, Vol. 73, pp. 247-256, 2017.

[12] D. T. Ha, C. N. Giap, P. T. T. Loan, and N.T. L. Huong, "An Empirical Study for Student Academic Performance Prediction Using Machine Learning Techniques", International Journal of Computer Science and Information Security, Vol. 18, No. 3, pp. 21-28, 2020.

[13] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition", Expert Syst. Appl., Vol. 41, No. 2, pp. 321-330, 2014.

[14] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques", Int. J. Mod. Educ. Comput. Sci., Vol. 8, No. 11, p. 36, 2016.

[15] R. Ghorbani, and R. Ghousi, "Comparing Different Resampling Methods in Predicting Student's Performance Using Machine Learning Techniques", IEEE Access, Vol. 8, pp. 67899-67911, 2020.

[16] A. Ghavidel, R. Ghousi, and A. Atashi, "An ensemble data mining approach to discover medical patterns and provide a system to predict the mortality in the ICU of cardiac surgery based on stacking machine learning method", Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, pp. 1-11, 2022.

[17] A. Desiani, S. Yahdin, and A. Kartikasari, "Handling the imbalanced data with missing value elimination SMOTE in the classification of the relevance education background with graduates employment", International Journal of Artificial Intelligence, Vol. 10, No. 2, pp. 346-354, 2021.

[18] C. W. Teoh, S. B. Ho, K. S. Dollmat, and C. H. Tan, "Ensemble-Learning Techniques for Predicting Student Performance on Video-Based Learning", International Journal of Information and Education Technology, Vol. 12, No. 8, pp. 741-745, 2022.

[19] A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications", 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, pp. 1200-1205, 2015.

[20] H. Li, and J. Sun, "Forecasting business failure: The use of nearest-neighbor, support vector and correcting imbalanced samples – evidence from the Chinese hotel industry", Tourism Management, Vol. 33, No. 3, pp. 622-634, 2012.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", Journal of Artificial Intelligence Reserach, Vol. 16, pp. 341-378, 2002.

[22] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new oversampling method in imbalanced data sets learning", Proc. Int. Conf. Intell. Comput. Berlin, Germany: Springer, pp. 878-887, 2005.

[23] Y. Tang, Y. Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification ", IEEE Trans. Syst., Man, Cybern, B. Cybern., Vol. 39, No. 1, pp. 281-288, 2009.

[24] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", IEEE World Congress on Computational Intelligence, pp. 1322-1328, 2008.

[25] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An Introduction to decision tree modeling", Journal of Chemometrics: A Journal of the Chemometrics Society, Vol. 18, No. 6, pp. 275-285, 2004.

[26] I. Rish, "An empirical study of the naive Bayes classifier", IJCAI 2001 workshop on empirical methods in artificial intelligence, Vol. 3, No. 2, pp. 41-46, 2001.

[27] D. Kurniadi, E. Abdurachman, H. L. H. S. Warnars, and W. Suparta, "The prediction of scholarship recipients in higher education using k-Nearest Neighbor algorithm", IOP conference series: material science and engineering, Vol. 434, No. 1, 2018.

[28] D. A. Pisner, and D. M.Schnyer, "Support vector machine in Machine learning", Machine learning Academic Press, pp. 101-121, 2020.

[29] S. F. Costa, and M. M. Diniz, "Application of logistic regression to predict the failure of students in subject of a mathematics undergraduate course", Education and Information Technology, pp. 1-7, 2022.

[30] S. Jeganathan, A. R. Lakshminarayan, and N. Ramchandran, "Predicting Academic Performance of Immigrant Students Using XGBoost Regressor", International Journal of Information Technology and Web Engineering, Vol. 17, No. 1, pp. 1-19, 2022.

[31] M. Ashraf, M. Zaman, and M. Ahmed, "Using Ensemble StackingC Method and Base Classifiers to Ameliorate Prediction Accuracy of Pedagogical Data", Procedia Computer Science, Vol. 132, pp. 1021-1040, 2018.

[32] M. Ashraf, M. Zaman, and M. Ahmed, "An intelligent prediction system for educational data mining based on ensemble and filtering approaches", Procedia Computer Science, Vol. 167, pp. 1471-1483, 2020.

[33] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining", Knowledge-Based Systems, vol. 200, pp. 1-16, 2020.

[34] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm", Interactive Learning ironments, pp. 1-20, 2021.

[35] M. Yagci, "Educational data mining: Prediction of students' academic performance using machine learning algorithms", smart learning environments, Vol. 9, No. 1, pp. 1-19, 2022.

[36] M. Ragab, A. M. K. A. Aal, A. O. Jifri, and N. F. Omran, "Enhancement of Predicting Students Performance Model Using Ensemble Approaches and Educational Data Mining Techniques", Wireless Communications and Mobile Computing, Vol. 2021, pp. 1-9, 2021.

[37] I. Nirmala, H. Wijayanto, and K. A. Notodiputro, "Prediction of Undergraduate Student's Study Completion Status Using MissForest Imputation in Random Forest and XGBoost Models", ComTech: Computer, Mathematics and Engineering Applications, Vol. 13, No. 1, pp. 53-62, 2022.

[38] S. Begum, and S. S. Padmannavar, "Genetically Optimized Ensemble Classifiers for Multiclass Student Performance Prediction", International Journal of Intelligent Engineering & Systems, Vol. 15, No. 2, pp. 316-328, 2022.

AI-based Academic Advising Framework: A Knowledge Management Perspective

Ghazala Bilquise[1]
Computer Information Science Department
Higher Colleges of Technology
Dubai, UAE

Khaled Shaalan[2]
Informatics Department
The British University in Dubai
Dubai, UAE

***Abstract*—Academic advising has become a critical factor of students' success as universities offer a variety of programs and courses in their curriculum. It is a student-centered initiative that fosters a student's involvement with the institution by supporting students in their academic progression and career goals. Managing the knowledge involved in the advising process is crucial to ensure that the knowledge is available to those who need it and that it is used effectively to make good advising decisions that impact student persistence and success. The use of AI-based tools strengthens the advising process by reducing the workload of advisors and providing better decision support tools to improve the advising practice. This study explores the challenges associated with the current advising system from a knowledge management perspective and proposes an integrated AI-based framework to tackle the main advising tasks.**

***Keywords*—Knowledge management; artificial intelligence; academic advising; rule-based expert system; machine learning; chatbot; conversational agent**

I. INTRODUCTION

Student retention and persistence are the most critical objectives of Higher Education Institutions (HEIs) as they are striving to meet the demands of the global economy. Graduating students on time is not just a measure of student and institutional success but also has a positive impact on the economy and society at large. In the United States, one out of three students does not progress from freshman to sophomore year, while in Australia nearly 30% of students do not graduate with a degree [1]. In UAE the rates are similar, with nearly 25-30% of students dropping out from a degree program [2]–[4]. With the astounding rate of university dropouts worldwide, academic institutions are striving hard to develop initiatives that mitigate the early leavers and provide the necessary support to students for on-time graduation [2].

Academic advising has been widely accepted as a vital strategy to tackle the problem of persistence and retention [5]–[7]. Advising is an essential process in academic institutions for engaging, supporting, and guiding students throughout their academic tenure. Tinto's prominent study [8] on the theoretical framework of retention states that students' engagement within the institution has a direct impact on reduced attrition rates. A broad definition of academic advising is provided by [9], who state that advising is the process ensuring student success through various interactions and between a student and members of the academic institution. Although there are several facets of academic advising, the main objective of advising is to effectively manage a student's journey to ensure academic success.

The process of advising encompasses several tasks such as ensuring students are informed about the institutional policies, courses, and program requirements and that they enroll for courses according to their degree plan. Furthermore, advising ascertains that students follow a customized learning track based on their academic progression [10]. Academic advising also offers extra support and guidance to students who need it the most [6], such as the students on probation or at risk of dropping out or failing a course.

Knowledge management (KM) is an integral part of academic advising. The process of advising involves the use of tacit and explicit knowledge to guide and support students throughout their academic life. Academic advisors assist students in various tasks such as selecting ideal courses, supporting at-risk students, providing necessary information that is vital to the student's successful integration with college life. Moreover, academic advisors also utilize their knowledge in solving issues that students face in achieving their academic goals. At the institutional level, the knowledge of the advisors must be captured, stored, and shared to ease the process of advising for new advisors as well as to retain knowledge within the institution. To this effect, technology provides an efficient means of disseminating knowledge among institutional members.

In the current age of digital transformation, Artificial Intelligence (AI) offers a promising avenue to effectively support the advisory process by providing benefits that are otherwise not attainable using a traditional advisory system. AI-based systems can automate the task of identifying students at risk, recommending courses, and answering student queries. These systems have the potential to not only reduce the workload of advisors but also enhance the services provided to students and support their academic progression [10]. Although there is a vast amount of research on supporting students using AI, there is no study that has investigated a comprehensive AI solution that tackles all the challenges of the traditional advising process.

The purpose of this study is to investigate the limitations of academic advising under the lens of KM and propose an AI-based solution for an academic institution based in the UAE. The study explores the problems of the advising system currently in place and offers a holistic solution based on AI

technologies that integrate with the current information system. The key requirements of the proposed solution are discussed with an implementation strategy.

The rest of the paper is organized as follows. Section II presents the background information of the higher education institution of this study and the advising process thereof. Section III highlights the problems of advising at the institution of study. Section IV reviews AI-based advising solutions in the existing literature. Section V discusses the proposed AI-based solution for the HEI of this study, and finally, the paper ends with a conclusion that provides a summary of the paper, limitations of the study, and further research avenues.

II. Background Information

The process of advising and the roles of the institutional members involved thereof may differ from one academic institution to another. This section describes the advising process followed at the institution of study and explains the roles of the academic advisor.

The academic institution of this study is one the largest higher education institution in the middle-east region. The institution offers six undergraduate programs of study and has an intake of nearly 500 students each term. After enrollment, students are provided credentials to access online resources such as the portal, emails, and the learning management system. Orientation sessions are held for the new students and an academic advisor is assigned.

Academic advising, at the institution of study, is a role assigned to every faculty member. Each faculty is assigned 25-30 advisees, who are students enrolled in the same program. The advising tasks and the advising process are consistent across all the programs. Therefore, this paper does not focus on any particular program of study, but rather the advising process as a whole.

An advisor's role encompasses three main tasks – creating a customized study plan for academic progression, providing guidance and support to answer queries and recommend opportunities for personal and career growth, and finally, monitoring academic progression and supporting students at risk.

First, an advisor liaises with each advisee to create a study plan by recommending courses every semester. A good study plan ensures a smooth academic progression in the program of study. The advisor must select appropriate courses that best meet the academic requirements such as pre-requisites, minimum credits, specialization, and more. The advisor also prepares a graduation plan during the final year of an advisees study to ensure on-time graduation.

The second advisory task is to offer guidance for general academic queries. The advisor is the central contact point for advisees who need direction and support for any personal or academic. An advisor directs the student to support systems provided by the institution such as student services, academic tutorials, or answers their general queries about grades, volunteering hours, GPA requirements, work placement, and more. This type of advising strengthens the student's bond with the institution as they feel connected to their environment. The advisor also corresponds with the advisees to encourage them to participate in extracurricular opportunities, competitions, and programs related to their career and personal growth. Moreover, advisees often reach out to their advisors for general guidelines and information on policies and procedures. The close interaction of advisees with their advisor leads to enhanced satisfaction level with the institution and reduces attrition rate [8].

The third advising task is the most crucial one as it is directly related to student success in the academic journey. It involves a pre-emptive check to follow up on students' academic progression, especially the students who are struggling with their studies. The advisor identifies and provides support to students who are at risk. The support may involve arranging a meeting with the counselor or facilitating tutorial sessions through the academic success center, or more. This type of advising has a significant impact on student retention and persistence [11].

A. Knowledge Management and Academic Advising

Knowledge management (KM) activities are at the core of the academic advising process. Therefore it is essential to understand KM and its application within the various advisory tasks. As new faculty members, and thereby advisors, join the institution, and current advisors leave, it is crucial to ensure that knowledge is captured and stored adequately to prevent knowledge loss. This section describes the KM processes and mechanisms involved in the advising process at the institution of study.

KM processes are the methods used to create, share and utilize knowledge within an organization. Study [12] identified four main KM processes - knowledge discovery, knowledge capture, knowledge sharing, and knowledge application. These processes are encompassed in all the advisory tasks as described below.

Knowledge discovery is the process of acquiring knowledge from various sources to make decisions, solve problems or generate new knowledge. Advisors use various sources of information such as the program structure, course requirements, and student's academic portfolio to build a customized plan for each student. They often brainstorm with other advisors and attend training to acquire knowledge related to this task.

Knowledge capture is the process of storing the acquired knowledge in a format that is readily available for those who need to access it. Advisors store the advising plans they have created in a student information system and share them with their advisees. However, a lot of the communication during this process is also captured in an unstructured format such as email, and in-person and phone conversations making it challenging to access and utilize this knowledge effectively in the future. Knowledge is also captured in the form of documentation of the policies and procedures of advising and is stored in the employee portal and communicated via email.

Knowledge sharing is the process of sharing tacit or explicit knowledge with other members of the institution. Advisors share their knowledge with advisees in the form of counseling, advice, and recommendations when performing advising tasks.

Moreover, advisors also share their best practices through informal and formal professional development sessions organized at the institution.

Knowledge application is the process of utilizing the knowledge to solve problems and perform tasks. Advisors use directions and routines to apply their knowledge based on the problem at hand and advisees' maturity level. For example, when dealing with new advisees, advisors direct the students on what courses to take in the first semester. As the advisee's maturity level increases, advisors guide the students by explaining how to choose courses and plan their studies.

Knowledge may be further subdivided into two main types –tacit and explicit knowledge. Tacit knowledge resides in the individual's mind in the form of experience, insights, and wisdom and is difficult to transfer, while explicit knowledge is documented and stored in a format that can be shared, understood, and applied. Reference [13] developed the spiral SECI model to explain how tacit and explicit knowledge interact with each other to create new knowledge. The model consists of four phases – Socialization, Externalization, Combination, and Internalization. Fig. 1 shows the tasks of advising in each phase of the SECI model.

The socialization phase occurs when advisors mentor the advisees using face or online meetings, share best practices with their colleagues, and attend professional development sessions to understand new technologies or requirements for advising. Socialization facilitates knowledge discovery and knowledge sharing processes. In this phase, tacit knowledge is used, which is largely based on experience, and advisor intuition.

The externalization phase occurs when tacit knowledge is converted into documented form. In the advising process, the registration department publishes policies, manuals, guidelines. The admissions department provides documentation on the program structure and student academic performance. The advisors utilize this information for effective advising. Although the general knowledge of advising is captured in the documentation, the specific knowledge that the advisors possess when dealing with various cases is lost, as the advisors are not required to externalize their knowledge on advising cases they have dealt with.

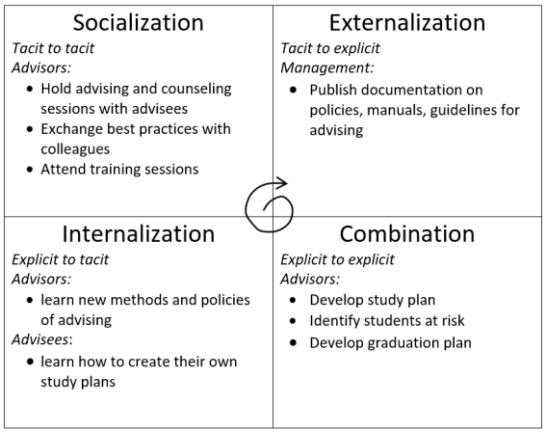


Fig. 1. SECI Model of Advising Tasks.

In the combination phase, the advisors integrate the explicit knowledge from various documentations to develop a customized study plan for every student and a graduation plan for final year students. The advisors also use the information to identify students at risk. Finally, in the internalization phase, the advisors learn new policies, methods, and systems essential for effective advising and students internalize the knowledge shared with them to create their own study plans.

This section provided an introduction to the paper, along with background information on the organization of the study and the role of KM in the advising process.

III. Problem Identification and Analysis

A. A KM Perspective of the Advising Challenges

People, processes, and technology are the three main interdependent elements of KM activities in an organization. The systematic integration of these three elements is essential to effectively implement KM practices in the organization [14], [15]. The current advising process at the institution of study has several limitations with regards to all these elements that hinder the integration of successful KM practices for the advising process. This section discusses the main challenges of the advising process.

1) People: People are the most essential element of KM practices in an institution as they are the possessors of knowledge. The people involved in the advising process are the advisors, students, and management staff. One of the main challenges is that the advisor does not have time to provide personalized advising to each advisee. The advisor's workload, of 20 teaching hours per week and involvement in research activities and other several administrative tasks, does not leave sufficient time for personalized interaction with 25 to 30 advisees. This has an impact on several tasks such as supporting students who are at risk, and maintaining a good level of communication and interaction with the advisees. The academic performance of a student is highly impacted by the quality of advising [9].

Another challenge is that new advisors do not have sufficient knowledge about the advising process to effectively advise students. Advisors must be aware of institutional policies, program structures, and academic requirements. Moreover, new faculty do not have the experience that advisors accumulated over the years. An inexperienced advisor does not have sufficient expertise to handle difficult cases such as students who have changed their programs and require course equivalency, or students on probation who need special attention in terms of planning courses. Erroneous advice in such cases may lead to a student repeating courses or taking courses that will not improve the GPA of a student on probation.

Communication between individuals in the advising process is vital to effective advising. Students are often shy to approach their advisors for queries as they do not know them personally. On the other hand, it is also observed that student queries are often repetitive relating to institutional policies and procedures such as registration times, applying for missed

assessment, following up registration, etc. Their general queries and concerns often go unaddressed, which in turn influences the student satisfaction level and integration at the institution. Moreover, advisors do not have the time to get to know each advisee personally. The lack of timely communication and interaction between advisee and advisor is a common challenge faced at the institution as it influences knowledge sharing negatively.

2) Process: The advising process includes tasks that are required for effective and efficient advising to capture and store the knowledge involved therein to make it available to those who need it. The SECI model for academic advising described in Section II, shows that the process of knowledge externalization is inadequate. Currently, the only form of documented explicit knowledge is provided by management in the form of manuals, policies, program requirements, and more. The knowledge accumulated by advisors over the years is not captured and shared in any formal way. This knowledge would be beneficial to both new advisors and current inexperienced ones. Furthermore, there is a risk of the knowledge being lost when a faculty members changes or leaves the job.

3) Technology: Technology acts as a supporting mechanism to facilitate the effective distribution and storage of knowledge to retain captured knowledge within the organization and make it available to individuals who need it [13]. At the HEI of study, several technologies are used to manage the information that is required for making informed decisions. For example, the advisee's academic performance data is stored in the banner system and available as reports for the advisor, the policies and procedures are stored in the SharePoint portal, and a degree audit system is used for managing and creating an advising plan. Moreover, email is also used for communicating new information, and requirements for advising. A lot of time and effort is spent in discovering knowledge from various sources for each student as these technologies are dispersed in different applications and not integrated.

Crucial information that is required by the advisor to track student progress and identify students at risk is currently not available during the semester. For example, the student's current semester's academic performance and attendance records are accessible by their teachers only. Due to this reason, advisors are unable to take pre-emptive measures at an early stage for an advisee who may be at risk of failure. Remedial actions are taken too late after the student has already failed the course.

The current system also does generate notifications to the advisor that are essential for their decision-making process. For instance, when an advisee drops the course or fails due to attendance, the advisor is not informed. This information is essential to modify the proposed study plan as it becomes a priority for the student to repeat the failed course to raise the GPA. At the end of the semester, when plans are updated based on student performance, the advisor has to manually check each student's academic record to update the student's plan.

TABLE I. ADVISING CHALLENGES AND ITS IMPACT

KM Element	Challenge	Impact
People	Lack of time	Advisors are unable to provide personalized advising and support to each advisee
	Lack of knowledge	Erroneous or inadequate advice by advisors, which may influence students' academic progression.
	Poor communication	Students find it challenging to integrate within the environment
Process	Lack of externalization	Organizational memory loss, advisors knowledge is not retained and shared in a formal way
Technology	Lack of structure	Knowledge is not captured in a structured format. Often involves email communication
	Lack of integration	Information is dispersed, requires time and effort to get access various information sources for decision making
	Lack of information availability	Advisors cannot take pre-emptive decisions and support advisees at risk at an early stage
	Lack of notifications	Cannot provide support at an early stage. Time-consuming to check each advisees' academic progression at the end of the semester.

Table I summarizes the advising problems faced at the institution of students and its impact on the institution and its members. The challenges highlighted below are the cause of inefficiencies in the advising process.

A software system is crucial to addressing the challenges of academic advising described in the previous section. Technology has the potential to automate tasks, reduce advising errors, improve communication, and provide insights on students' progress. To this effect, Artificial Intelligence (AI) based technology solutions offer a promising avenue to address the challenges of the current advising process. The AI-based tools can automate the low-impact tasks to reduce the workload of advisors and provide insights for key tasks to support better decision-making [16]. Moreover, AI also has the potential to enhance students' experience through machine intelligence supported by human advisors [9].

Fig. 2 presents a visualization of the analysis of terms in research studies related to AI in higher education over the last two decades. The visualization, constructed using VOSViewer depicts the relationships between frequently occurring terms in the research papers in the form of a network diagram [17]. Three main clusters are evident in the diagram. The red cluster shows that data mining is the predominant study in HEIs, which has been investigated by a vast majority of authors. The blue cluster shows that the main techniques researched are machine learning algorithms. The studies on academic advising are very limited. In addition to this, the studies that have investigated academic advising have only examined one aspect of advising rather than providing a comprehensive AI solution that tackles all the advising problems.

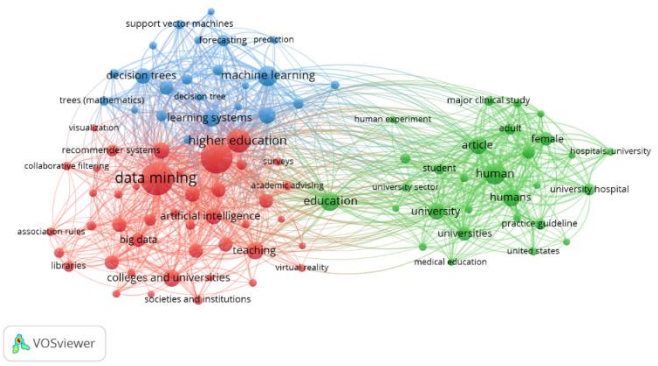


Fig. 2. VOSViewer Analysis of Terms.

This study proposes AI-based technologies that offer a comprehensive solution to make the advising process more effective and efficient. At the institution of study, the following three main tasks of advising have been identified that can be automated using AI-based advising tools to alleviate the underlying problems described in this section:

- Developing personalized study plans
- Early identification of students at risk
- Provide personalized assistance to students.

IV. Literature Review

This section provides a critical review of the use of AI-based solutions for advising in the existing literature. The review focuses on the three main advising tasks outlined in the paper. Several studies have investigated AI-based advising solution that merely examines one area of advising, while only a few have applied AI solutions to multiple advising tasks. Thus, this section reviews studies by categorizing them based on the type of AI solution proposed by reviewing studies that have focused on a single aspect of advising and those that have used AI-based solutions for more than one advising task.

A. AI-Based Solution for Study Plans

Numerous papers have researched the use of AI for creating study plans or recommending the ideal courses for students to maximize success. While some studies focus on rule-based systems, others are based on machine learning algorithms for recommending ideal courses to students that maximize their success.

A data-driven model is used by [18] to create a predictive analytics tool that supports academic advisors in advising decisions based on insights from historical data. They use a web application with a rich dashboard interface to display the chances of student success in the selected courses along with the details of the prediction. A multilevel clustering algorithm is used to predict the success rate in each selected course based on previous students' academic performance data such as course grades and the number of courses registered in a semester. The authors used a comparative study to verify their system in two universities. The participants of the study, experienced and inexperienced academic advisors, used traditional methods and the predictive analytical tool to perform advising for several cases. The results of the study showed that advisors explored more course options when using the AI-based tool to develop a suitable study plan with lower failure risk for each student. The main limitation of the study is that it relies on the academic advisor to select the courses for the student and the system merely provides a success rate of the selected courses. The system relies on the adivsors knowledge of institutional, course, and program requirements for selecting ideal courses.

Reference [19] proposed an intelligent advising system that assists students in course registration. The system is intended to be used by students without the need of a faculty advisor. Students are recommended courses based on their current performance, known preferences, historical data, and academic policies. The proposed advising system integrates with the current information system for the academic data required for predictive analytics. The system uses association rule mining to explore patterns in the academic dataset to identify the group of courses that should ideally be taken together. A rule-based expert system is used to assign a priority score to the courses based on academic policies and factors such as student GPA and nature of the course (prerequisite course, core or elective course etc.), course grade, and more. Finally, a recommendation algorithm is used to suggest courses to the students. A limitation of the study was that it considered student preference as a major factor in the recommendation model, but failed to describe the features used to determine students' interests. Moreover, the proposed model was not evaluated for the quality and accuracy of the recommendations.

Another model proposed by [20] recommends courses to university students based on personal traits and academic data. Student personal characteristics include features like gender, age, knowledge level, learning style, the term of study, and performance. The academic data consists of features such as courses, credit hours, semester of study. The proposed model uses a knowledge-based model to assign weights to selected courses based on the students' performance. The study does not apply institutional policies, course, and program requirements when recommending courses.

Study [21] designed an interactive system to recommend suitable courses to university students based on their interest and popularity of the course. The recommendation is based on historical enrollment data, course descriptions, topic, instructor, and time of study. A student searches for the offered course using keywords and may filter the popular course recommendations by providing preferences such as time of class, topic of interest, and more. The system has several limitations. First of all, it does not integrate with the current information system. Recommendations are not based on students' academic history or performance. Second, the system is only suitable for universities with a flexible curriculum where a student is free to explore and take various courses across different departments.

Both [22] and [23] developed a rule-based expert system that recommends courses for university students. The expert system rules are based on course pre-requisite requirements, year of study, and course eligibility. The system provides a rationale for each recommendation. The study [23] does not integrate the system with the data stored in the student

information system. The student is required to provide the courses they have completed, their current GPA, and their major as input to the system. Moreover, failed courses are not taken into account when making the recommendation. The research [22] integrated the expert system with data extracted from the institutional database. However, the system does not prioritize the recommended courses according to the importance of registration in the following semester.

B. AI-Based Solution to Identify Students at Risk

The use of machine learning algorithms to develop automated intervention systems that integrate with a learning management system (LMS) has been investigated by several researchers. Students' engagement in the online environment and their current academic performance can be used to predict course outcomes at an early stage [24]. Furthermore, studies have also shown that timely interventions and support for low-performing students are effective to help them manage their study patterns [25].

Study [26] used machine learning and deep learning algorithms for the early identification of students at risk using data collected from an online learning platform. The study predicts student failure at various stages of course completion based on the student demographic data, performance, and engagement data using click patterns. The study shows that the random forest algorithm performed the best with up to 92% precision, recall and accuracy. The study further recommends intervention strategies at the various course completion stages based on prediction outcomes by sending messages of encouragement, recommendation, or fear to students at risk. The study does not consider the involvement of the advisor or instructor in supporting underperforming students. Students who are at risk may not possess the mental or emotional capability to comprehend the motivational intervention messages. The involvement of the advisor is essential to determine the support a student may require to improve his performance.

Reference [27] used deep learning to identify students at risk of drop-out at an early stage in an online course. The study uses click patterns, discussion, and quiz scores to create prediction models using SVM, KNN, decision tree, and deep learning algorithm to predict student dropout at a weekly rate. The deep learning algorithm performed best with an average AUC (area under the curve) rate of 96%. The study further went on to suggest intervention strategies based on the probability of course dropout, such as varying levels of support by the instructor.

Both the studies [26], [27] were based on Massive Online Open Course (MOOC) dataset, that have a large number of enrollments and thus a huge dataset that is required for deep learning. On the contrary, enrollment in degree programs will not have the same dataset size collected from the virtual learning environment. Moreover, the models were not tested on different sized datasets for the generalization of results.

Study [28] used clickstream data collected from an ebook interaction log, along with student performance at various stages in the course to predict student performance. Comparisons of prediction accuracies during various weeks of the course showed that the earliest reasonable accuracy, of 79%, is achievable as early as week 3. The study is based on the assumption that the ebook is the main resource used by all students in the online learning environment. Furthermore, the study did not utilize data from the existing information system to generate the predictive model.

Eight machine learning algorithms were used by [29] to determine the optimal time during a semester-length course to predict student grades. The study uses student demographic data, academic data, weekly assessment scores, and LMS interaction data to create a prediction model. Weekly predictions revealed that the earliest reasonable prediction rate is achievable by week six to support early intervention for poorly performing students. The study relies on continuous weekly assessments for predictions, which is not applicable in most courses. Moreover, the study integrated LMS data with student admission and academic background data but did not consider attendance as a feature for prediction.

C. AI-Based Solution for Digital Assistance

With today's technological advancement students are constantly in need of information for their daily tasks and academic progression. Providing adequate channels for student communication is vital to help students integrate with their environment and feel connected and enhance student satisfaction. Students often have queries about the institutional and academic policies and procedures, academic progression, activities, and more. In reality, the student services team and the academic advisors are usually overwhelmed with such a large number of queries that they are not able to provide instant responses. As a result students' disconnection and dissatisfaction with the institution increases.

Chatbot systems have the potential of providing students with the information they need by answering their queries in a conversational style. They provide 24/7 service, unlike human advisors. Despite the numerous benefits of chatbots in improving levels of service, the use of chatbots in HEI for advising is very limited [30]. This section reviews the AI-based solutions that have used a chatbot system for improving communication and answering student queries.

Study [31] designed a rule-based expert system that answers students' queries on institutional policies and guidelines to familiarize students with the environment. The digital assistant, built with CLIPS and JAVA, uses both forward and backward chaining and is based on inference rules. The knowledge base for the expert system was gathered from the website, student feedback, and experts in the institution. User queries were classified into four categories – yes/no, what, where, and when questions. The automated virtual assistant was tested completeness and correctness using 70 participants and resulted in an accuracy of 99%. The main limitation of the study is that the chatbot system does not support conversational AI. The question type has to be selected from a predefined list. Interaction with natural language processing would be more intuitive and adaptive for end users.

An intelligent academic advisor using a DeepQA system built was built by [32] using IBM Watson. The system was used to answer queries from potential, new, and, current

students as well as faculty members pertaining to academic advising in a business school. The intelligent system was initially populated with a database of nearly 300 questions and answers, and other information extracted from FAQs, syllabus, and more. Moreover, the intelligent system has an engine to learn and increase its knowledge base. The chatbot does not provide personalized feedback to students.

Reference [33] used a conversational agent to support administrative tasks of recruiting students into degree programs. The AI system matched student skills to the program requirements by asking questions and using keywords from their answer to select suitable programs. The admin staff also utilize the system to query information about shortlisted candidates. The system was not designed mainly for administrative purposes and not for advising.

Study [34] used a chatbot to ease the process of selecting elective courses for a degree program in computing. The chatbot uses natural language to answer queries related to the courses, provide peer reviews about the courses, analysis of choices, and provide a personalized recommendation based on the student record. The chatbot had a very specific use and did not provide advising in other matters of academic life.

Reference [35] used a chatbot application, developed using the IBM Watson API, to provide support to students struggling in programming. The chatbot not only provides support for programming related queries, but also for personal issues such as depression, suicidal thoughts, etc. It directed students to the appropriate department call center for their issues or calls the ambulance based on the severity of the case. The chatbot was designed to detect student frustration while studying programming. It did not provide general assistance in other college related matters.

D. AI for Multiple Advising Tasks

Some studies investigated an AI-based solution for more than one aspect of advising. Latorre-Navarro (2014) developed an AI-based solution that answers student queries and recommends courses. A conversational agent was used to answer questions on a wide range of topics related to academic policies, procedures, and services. The authors also used an expert system that guides students to create their study plan and sends it to the advisor for approval. The main limitation of the system is that it is not integrated with the information system. Students are required to provide their academic progress such as current courses, failed courses, and completed courses. An error in providing this information could result in an incorrect plan.

An intelligent web-based advising system that supports effective advising was developed by [36]. The system is designed to be used by both advisors and students. A rule-based expert system is integrated with the current information system to extract a student's academic record and create a study plan for the following semester and view the graduation status. The system also answers basic queries related to institutional policies. Advisors can also use the system to view their advisees' profiles, and get access to all the advising documentation integrated in a single location. Notifications are sent to the advisor when there is an update to a policy, ensuring that all advising decisions are accurate. A limitation of the study is that it does answer any personalized queries or send reminders notifications to students.

Though both studies [16], [36] tackled more than one advising problem by leveraging AI-based technologies, yet they do provide a comprehensive solution. The studies did not investigate one of the main tasks of advising, which is to identify and support low-performing students with early intervention strategies.

An overview of the literature shows that, to the best of the author's knowledge, no study exists that provides a holistic advising solution using AI technologies to addresses all the challenges of advising faced at an academic institution. Hence the purpose of this study is to fill the gap in this area and recommend a comprehensive AI solution for the institution of study.

V. Identification and Evaluation of Alternate Solutions

This section discusses three AI-based solutions proposed for the institution of study – (1) AI-based solution for creating study plans (2) AI-based solution for identifying students at risk of failing a course at an early stage, and (3) AI-based solution for personalized digital assistance. All solutions are integrated with the institutional database to provide personalized information to the students to support their academic progression. The study proposes the use of a rule-based expert system to create ideal study plans, a machine learning model to identify students at risk, and a chatbot system to provide personalized digital assistance. Fig. 3 shows an overview of the proposed solution.

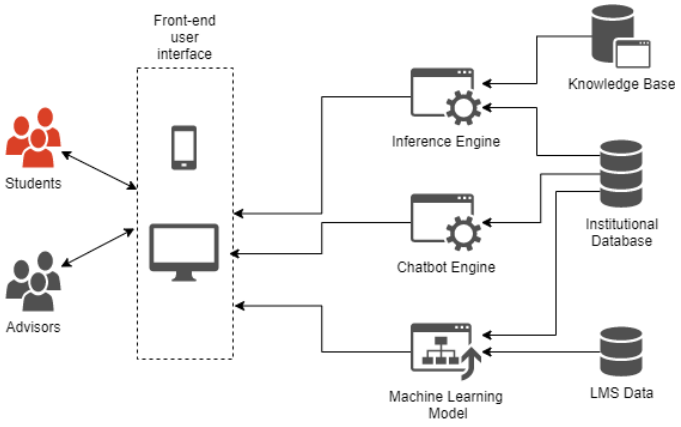


Fig. 3. Overview of Proposed Solution.

A. AI-Based Solution for Creating Study Plans

Studies have investigated the use of machine learning for recommending courses to students [21] as well as including features personal traits [20], and student preferences [19]. While this type of model is suitable for online courses with a large number of enrollments and course choices, this model does not work for the institution of study. The courses offered at the current institution are based on a program requirement that has a predefined number of courses with a few electives. The courses that must be taken according to the ideal semester plan, obeying the rules such as course sequence in the program

structure, minimum and maximum required credits (with the program area), catalog term, academic progression of the student, and more.

To this effect, this study proposes a rule-based expert system that captures the knowledge of domain experts to create a knowledge base. The expert system utilizes student data from the institutional database and applies registration rules and policies to recommend a list of ideal courses that maximize the chance of students graduating on time. Fig. 4 shows the architecture of the expert system.

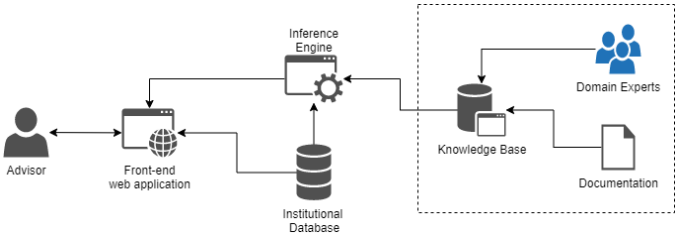


Fig. 4. Rule-Based Expert System.

The knowledge engineer will acquire the knowledge from domain experts such as the registration staff, and expert advisors to build a knowledge base. All the policies and registration requirements are also encoded as rules with the system. The system uses student academic data, program data, and course data from the institutional database as initial facts to map the student course requirements against the program requirements and applies the rules to identify a list of ideal courses.

- Student Academic Data consists of the catalog term, the program of study, placement scores, list of all completed courses, failed courses, and credit hours completed.
- Program Data consists of the requirements of the program such as the total credits required in each area (core courses, elective courses, concentration courses, and general studies courses).
- Course Data consists of the credit hours of the course and the pre-requisite(s), co-requisite(s), and equivalent courses.

The inference engine applies the knowledge base rules to the student, program, and course data that are the initial facts in the working memory. The eligible courses are assigned a priority based on the importance of completing that course in the following semester. For example, a higher priority is assigned to a course in which the student previously failed, or a course that is a pre-requisite of other courses in the following semester. Finally, a web-based interface is used to present the study plan to the advisor, in order of priority. The advisor analyzes the plan that is created by the system and makes any necessary modifications and advises the student accordingly.

B. AI-Based Solution for Identifying Students at Risk

Machine learning algorithms have been investigated in numerous studies to identify students at risk at an early stage during course progression. Some studies relied on LMS click patterns to predict low-performing students [26]–[28]. LMS interaction requires interacting with the course content online, which in turn generates a click pattern that can be analyzed for student engagement within the course. This model is not appropriate for the institution of the study, as most of the courses are face-to-face. Students mainly use the online environment to download course resources, attend online sessions, or submit assessments. Click patterns would not be an ideal indicator of student engagement especially when the student is using the course resources offline.

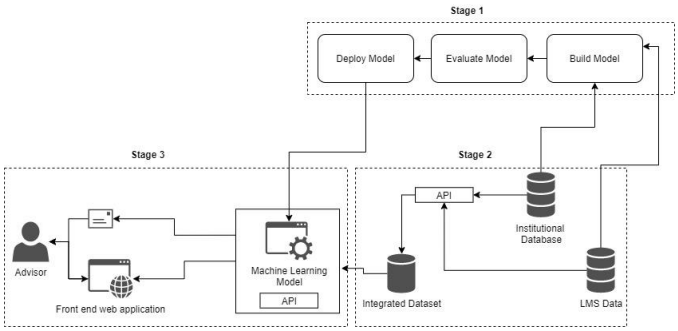


Fig. 5. Mahcine Learning Model for Identifying Students at Risk.

To this effect, this study proposes the use of course performance data and student academic history to predict the risk of failing a course. The proposed system will be used to low performing students as early as week six so early intervention strategies can be engaged. Howard et al. (2018) showed that reasonable machine learning performance can be achieved at week 5-6 to predict course attrition. Fig. 5 shows the architecture of the proposed advising system.

The system integrates LMS data and institutional data to build a machine learning model. The main elements of the system are explained in three stages:

1) Build and deploy the machine learning model: In this stage data historical is extracted from the institutional database and the LMS at the beginning of the academic year. The institutional database contains the following data:

- Students enrollment records that include the high school score, IELTS score, placement test scores, gender and status (working or not), and other profiling information
- Academic data such as program of study, credit hours completed, credit hours registered, courses completed, overall GPA, and attendance record.

LMS Data contains the coursework assessment data. Coursework assessments are usually conducted at regular intervals – week 6, week 12, week 15. The final assessment is scheduled on week 16.

In stage 1, the machine learning model is developed using historical data of the last 4-5 years. The data will be pre-processed and used for training and testing multiple machine learning algorithms. Several machine learning algorithms will be used and evaluated to determine the best performing algorithm suited for the data provided. At the end of this stage, the machine learning model will be deployed for use with the current records. This stage will be repeated once every academic year for monitoring and tuning the model to generate a new model based on new data acquired in the previous year.

2) Integrate LMS and institutional data: In this stage, the current semesters academic record is extracted from both the institutional and LMS and integrated into a dataset. It is recommended that the extraction takes place at week 6, week 12, and week 15. Based on the findings of previous researchers [28], [29], it is expected that good prediction rates of at-risk students are achievable by week 6.

3) Apply the machine learning model to generate a prediction: In this stage, the deployed machine learning model is applied to the extracted data of the current semester to determine students at risk of failure. A web application is used as a front-end interface for advisors to view the risk profile of their advisees. The system also sends notifications about advisees that require immediate attention.

C. AI-Based Digital Assistant

A conversational AI-based solution provides timely response to students' academic queries and improving the students' experience. Several studies have investigated the use of chatbots in an educational setting, however, the main purpose of the system was either administrative, such as recruiting students [33], or recommending courses [34]. Moreover, the studies that used chatbots for answering student queries [31], [32] did not integrate it with student's academic record to provide personalized information. None of the chatbots proposed in the reviewed studies send push notifications to students.

This study recommends the use of a conversational AI chatbot that integrates with the institutional database containing student's academic history, registration schedules, program requirements, course requirements, and a knowledge base of frequently asked questions. Unlike a human advisor, the chatbot will be available 24/7 to respond to various student queries. It will respond to general queries and personalized queries using the student's academic data. Furthermore, the chatbot will also initiate reminders to the student about general information such as upcoming deadlines, and initiate personalized reminders such as high absences rates to ensure that the student does not miss more classes. The notifications may also be personalized to the student's interest such as sports, clubs, and more.

Examples of general student queries are:

- When is the deadline for dropping a course?
- What is the pre-requisite for CIS 2203?
- How can I change my program?

Examples of personalized student queries are:

- What is my CGPA?
- How many volunteering hours have I completed so far?
- How many absences do I have?
- Who is my advisor?

Example of general notification:

- Add and drop period ends on Sunday, 10th October

Example of personalized notification:

- You have reached 7% absence in the advanced programming course.

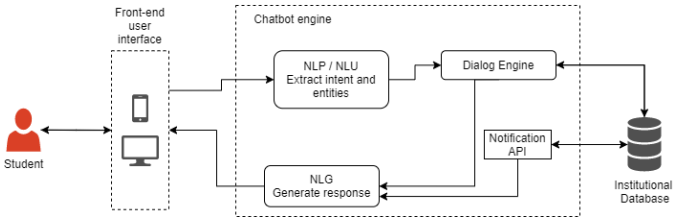


Fig. 6. Advising Chatbot Architecture.

Fig. 6 shows the proposed chatbot architecture. A brief explanation of the architecture is given below:

- The student writes a query in natural language using the chatbot client interface.
- The chatbot backend processes the query using a NLP (Natural Language Processing) engine, which converts the written text to structured data.
- The NLU (Natural Language Understanding) engine then extracts the intent and entities from the given structured query.
- Based on the processed query, the dialog engine of the chatbot retrieves the data from the institutional database and presents the response to the NLG (Natural Language Generator).
- The NLG processes the structured response into natural language and presents it to the student via the front-end interface
- The chatbot engine also contains an API, linked to a scheduler, which retrieves data from the database to send timely reminders and notifications to the students.

The proposed advising system reduces the workload of advisors by automating repetitive and mundane tasks. Advisors can spend their time getting to know their advisees and supporting them in their personal and career growth. The benefits of the proposed AI-based solution are summarized in Table II.

TABLE II. ADVISING CHALLENGES AND ITS IMPACT

AI-Based Solution	Benefits
Rule-based expert system for creating study plans	• Integrates with the existing information system to use accurate data of current student records, academic history, registration, and academic policies. • Minimizes/eliminates erroneous decisions of new or inexperienced advisors • Reduces workload of advisors
Machine learning model for identifying students at risk	• Provides early identification of students at risk before it is too late. • Support students with early intervention strategies that can possibly alleviate the risk of failure
Conversational AI chatbots for digital assistance	• Constant availability of support to students enhances student experience and increases loyalty towards the institution. • Provides equal opportunity for all students to ask questions, at any time 24/7 • Access to personalized assistant • Students are encouraged to stay on track with nudges from the system such as reminders

VI. CONCLUSION

Academic advising is a vital function in HEIs for providing guidance and support to students throughout their academic tenure [5]. Effective advising not only has a significant impact on students' academic performance but also has a positive influence on the overall academic experience [37] contributing to academic retention and persistence. The imperative nature of academic advising makes it crucial for institutions to invest in tools that support advisors in managing advising tasks effectively. This study explored the practice of academic advising at an academic institution based in the UAE. The study investigated the advisory process and the limitations under the umbrella of Knowledge Management. Finally, AI-based technologies are proposed to provide a comprehensive solution that automates advising tasks. AI-based systems can guide students through their journey with little intervention from the advisors, thus reducing the workload of advisors from menial tasks to focus their effort on key advising tasks such as development advice and career planning. Moreover, AI-based advisory systems may be personalized for individual student need and provides an equal opportunity for all students to access the information and service that they need.

The problems of academic advising are highlighted in this study from the perspective of the three KM elements - people, processes, and technology. Students and advisors are the people involved in the advisory process. The ratio of advisor to advisees makes it challenging for advisors to provide personalized guidance to each advisee. Moreover, an advisor's inexperience or lack of knowledge may lead to erroneous advice. Student queries often go unanswered leading to dissatisfaction and frustration. The advisory process involves creating study plans for students, dealing with issues, guiding and counseling students. An experienced advisor's knowledge is currently not captured in any formal way and may lead to organizational memory loss when the advisor leaves the institution. The current technologies at the institution are inadequate in supporting all the KM processes involved in advising effectively. The information is dispersed in different systems making it inefficient to look up for each student. Advisor's time and effort are consumed in analyzing student data, to create study plans. Furthermore, advisors need to set meetings with advisees, send reminders, and follow up on failed courses. Moreover, the current system does not provide insights to identify low-performing students so pre-emptive measures may be taken to manage the course of their studies.

The study proposes three AI-based systems as a comprehensive solution to alleviate all the problems associated with the current advising process – (1) Rule-based expert system for recommending courses and developing study plans for the following semester (2) Machine learning algorithm for identifying students at risk of failing a course at an early stage, and (3) conversational AI chatbots to provide personalized digital assistance to the student. All three systems integrate with the data in the current information system to provide personalized support and guidance to students and advisors. The systems are promising in terms of reducing the advisor's workload and improving student satisfaction with the institution, leading to student retention and persistence as an overall goal.

A. Limitations and Future Research

This study provides a framework for leveraging AI technologies in academic advising and focuses only on the three main tasks performed by an academic advisor. An avenue of future research is to investigate the implementation of the three systems as a prototype and a proof of concept. The systems must be verified for accuracy and quality of results.

One of the aspects of advising that is not considered in this study is the guidance provided to students during the enrollment stage to choose a program of study. This type of advising is done by the admission department and is crucial as most students are undecided about their career pathways when they enroll. Furthermore, studies have shown that students often receive inadequate guidance to make the right program choice, which in turn leads to changing programs during their studies [38], thus delaying graduation. In some cases, it may also lead to dropping out due to lack of interest or inability to cope with the program requirements. Machine learning algorithms may be investigated for recommending ideal programs to students that maximize their chances of success.

REFERENCES

[1] K. MacGregor, "Access, retention and student success – A world of difference," University World News, 2020. https://www.universityworldnews.com/post.php?story=20200904081106566 (accessed Nov. 03, 2021).

[2] A. Naidoo, "Early warning software helps prevent dropouts in UAE | Education – Gulf News," Gulf News, 2010. https://gulfnews.com/uae/education/early-warning-software-helps-prevent-dropouts-in-uae-1.729186 (accessed Nov. 15, 2021).

[3] H. M. Elmehdi, E. Z. Dalah, A. Bukhatir, and A. M. Ibrahem, "Retention at the University of Sharjah: Factors and Strategies," in 2020 Advances in Science and Engineering Technology International Conferences (ASET), 2020, pp. 1–6.

[4] UAEU, "Retention and Graduation Rates," 2020. https://www.uaeu.ac.ae/en/about/ss@uaeu/retention-and-graduation-rates.shtml (accessed Nov. 15, 2021).

[5] S. Campbell and C. Nutt, "Academic Advising in the New Global Century: Supporting Student Engagement and Learning Outcomes Achievement," Peer Rev., vol. 10, no. 1, p. 4, 2008.

[6] J. K. Drake, "The Role of Academic Advising in Student Retention and Persistence," About Campus Enrich. Student Learn. Exp., vol. 16, no. 3, pp. 8–12, 2011, doi: 10.1002/abc.20062.

[7] T. Fricker, "The Relationship between Academic Advising and Student Success in Canadian Colleges: A Review of the Literature.," Coll. Q., vol. 18, no. 4, p. n4, 2015.

[8] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," Rev. Educ. Res., vol. 45, no. 1, pp. 89–125, 1975.

[9] A. Assiri, A. A. M. Al-Ghamdi, and H. Brdesee, "From traditional to intelligent academic advising: A systematic literature review of e-academic advising," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 4, pp. 507–517, 2020, doi: 10.14569/IJACSA.2020.0110467.

[10] O. Iatrellis, A. Kameas, and P. Fitsilis, "Academic advising systems: A systematic literature review of empirical evidence," Educ. Sci., vol. 7, no. 4, 2017, doi: 10.3390/educsci7040090.

[11] A. Y. Noaman and F. F. Ahmed, "A New Framework for e Academic Advising," Procedia Comput. Sci., vol. 65, no. Iccmit, pp. 358–367, 2015, doi: 10.1016/j.procs.2015.09.097.

[12] M. Alavi and D. E. Leidner, "Knowledge management and knowledge management systems: Conceptual foundations and research issues," MIS Q., pp. 107–136, 2001.

[13] I. Nonaka and H. Takeuchi, The knowledge-creating company: How Japanese companies create the dynamics of innovation. Oxford university press, 1995.

[14] Y. Butler, "Knowledge management—if only you knew what you knew," Aust. Libr. J., vol. 49, no. 1, pp. 31–43, 2000.

[15] M. Evans, K. Dalkir, and C. Bidian, "A holistic view of the knowledge life cycle: the knowledge management cycle (KMC) model," Electron. J. Knowl. Manag., vol. 12, no. 1, p. 47, 2015.

[16] E. M. Latorre-Navarro, An intelligent natural language conversational system for academic advising. University of Florida, 2014.

[17] N. J. Van Eck and L. Waltman, "VOSviewer manual," Leiden: Univeristeit Leiden, vol. 1, no. 1, pp. 1–53, 2013.

[18] F. Gutiérrez, K. Seipp, X. Ochoa, K. Chiluiza, T. De Laet, and K. Verbert, "LADA : A learning analytics dashboard for academic advising," Comput. Human Behav., vol. 107, no. December 2018, p. 105826, 2020, doi: 10.1016/j.chb.2018.12.004.

[19] L. Aynekulu and T. Boran, "An intelligent and personalized course advising model for higher educational institutes," Appl. Sci., vol. 2, no. 10, pp. 1–14, 2020, doi: 10.1007/s42452-020-03440-4.

[20] A. A. Al-Hunaiyyan, A. T. Bimba, and S. Alsharhan, "A Cognitive Knowledge-based Model for an Academic Adaptive e-Advising System," Interdiscip. J. Information, Knowledge, Manag., vol. 15, pp. 247–263, 2020.

[21] B. Ma, M. Lu, Y. Taniguchi, and S. Konomi, "CourseQ: the impact of visual and interactive course recommendation in university environments," Res. Pract. Technol. Enhanc. Learn., vol. 16, no. 1, 2021, doi: 10.1186/s41039-021-00167-7.

[22] G. M. S. Alfarsi, K. A. M. Omar, and M. J. Alsinani, "A Rule-Based System for Advising," J. Theor. Appl. Inf. Technol., vol. 95, no. 11, 2017.

[23] G. Engin, B. Aksoyer, M. Avdagic, and D. Bozanl, "Rule-based expert systems for supporting university students," 2nd Int. Conf. Inf. Technol. Quant. Manag., vol. 31, pp. 22–31, 2014, doi: 10.1016/j.procs.2014.05.241.

[24] C. J. Villagrá-Arnedo, F. J. Gallego-Durán, F. Llorens-Largo, P. Compañ-Rosique, R. Satorre-Cuerda, and R. Molina-Carmona, "Improving the expressiveness of black-box models for predicting student performance," Comput. Human Behav., vol. 72, pp. 621–631, 2017, doi: 10.1016/j.chb.2016.09.001.

[25] K. E. Arnold and M. D. Pistilli, "Learning Analytics to Increase Student Success," 2nd Int. Conf. Learn. Anal. Knowl., no. May, pp. 267–270, 2012.

[26] M. Adnan et al., "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models," IEEE Access, vol. 9, pp. 7519–7539, 2021, doi: 10.1109/ACCESS.2021.3049446.

[27] W. Xing and D. Du, "Dropout Prediction in MOOCs : Using Deep Learning for Personalized Intervention," J. Educ. Comput., vol. 57, no. 3, pp. 547–570, 2019, doi: 10.1177/0735633118757015.

[28] G. Akçapınar, M. N. Hasnine, R. Majumdar, B. Flanagan, and H. Ogata, "Developing an early-warning system for spotting at-risk students by using eBook interaction logs," Smart Learn. Environ., vol. 6, no. 1, 2019, doi: 10.1186/s40561-019-0083-4.

[29] E. Howard, M. Meehan, and A. Parnell, "Contrasting prediction methods for early warning systems at undergraduate level," Internet High. Educ., vol. 37, no. January, pp. 66–75, 2018, doi: 10.1016/j.iheduc.2018.02.001.

[30] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," Comput. Educ. Artif. Intell., vol. 2, p. 100033, 2021, doi: 10.1016/j.caeai.2021.100033.

[31] P. Lodhi, O. Mishra, S. Jain, and V. Bajaj, "StuA : An Intelligent Student Assistant," Big Data Open Educ., no. February, pp. 17–25, 2018, doi: 10.9781/ijimai.2018.02.008.

[32] C. Asakiewicz, E. A. Stohr, and S. Mahajan, "Building a Cognitive Application Using Watson DeepQA," IT Prof., vol. 19, no. 4, pp. 36–44, 2017.

[33] W. A. Elnozahy, G. A. El Khayat, L. Cheniti-Belcadhi, and B. Said, "Question Answering System to Support University Students' Orientation, Recruitment and Retention," Procedia Comput. Sci., vol. 164, pp. 56–63, 2019, doi: 10.1016/j.procs.2019.12.154.

[34] C. H. Chan, H. L. Lee, W. K. Lo, and A. K.-F. Lui, "Developing a Chatbot for College Student Programme Advisement," Proc. - 2018 Int. Symp. Educ. Technol. ISET 2018, pp. 52–56, 2018, doi: 10.1109/ISET.2018.00021.

[35] [35] M. Ismail and A. Ade-Ibijola, "Lecturer's Apprentice: A Chatbot for Assisting Novice Programmers," Proc. - 2019 Int. Multidiscip. Inf. Technol. Eng. Conf. IMITEC 2019, pp. 1–8, 2019, doi: 10.1109/IMITEC45504.2019.9015857.

[36] L. Keston and W. Goodridge, "AdviseMe: An Intelligent Web-Based Application for Academic Advising," Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 8, 2015, doi: 10.14569/ijacsa.2015.060831.

[37] A. D. Young-jones, T. D. Burt, S. Dixon, and M. J. Hawthorne, "Academic advising: does it really impact student success?," Qual. Assur. Educ., vol. 21, no. 1, pp. 7–19, 2013, doi: 10.1108/09684881311293034.

[38] M. S. Jaradat and M. B. Mustafa, "Academic advising and maintaining major: Is there a relation?," Soc. Sci., vol. 6, no. 4, 2017, doi: 10.3390/socsci6040151.

An Adaptation Layer for Hardware Restrictions of Quadruple-Level Cell Flash Memories

Se Jin Kwon
Department of AI Software
Kangwon National University
Samcheok, South Korea

***Abstract*—In recent years, major flash memory vendors have produced SSDs and fusion memories as substitution for hard disks. However, there has been a lack of studies on access restriction of QLC flash memory, since most researches have targeted small capacity flash memory. As a solution, we propose to implement an adaptation layer between the file system and FTL (Flash Translation Layer). Instead of immediately writing data given from file system to flash memory, the adaptation layer gathers and adjusts data in the unit of a page, and separates random data from sequential data. By implementing the adaptation layer, previous FTL algorithms can be fully applied on the QLC flash memory. According to our experiment, the adaptation layer forms smaller number of pages than the current data gathering algorithm.**

Keywords*—Cache storage; flash memory; SSD; nonvolatile memory

I. Introduction

The capacity of flash memory has been rapidly growing as it has been introduced as a new solution for substituting the hard disks. Current QLC flash memories such as SSDs and fusion memories contain pages that are four to eight times larger than file system's data sector [1]. Due to the large page size of QLC flash memories, it is required to re-access a page to write multiple file system's data sectors within a page. However, the number of partial programming (NOP) within a page is limited to only one to avoid program-disturb errors [2]. Therefore, QLC flash memory uses an internal buffer to gather data in a unit of a page before writing onto the flash memory.

Unfortunately current well-optimized FTLs do not contain data gathering algorithm for NOP restriction, since they are designed for small capacity flash memories [3]. In this paper, we are not concerned with developing efficient mapping algorithm, since the small capacity based FTL algorithms already give various solutions. Instead, we propose to implement an adaptation layer between file system and FTL. It enables the small capacity based FTL algorithms to be fully applied on the QLC flash memory. Instead of immediately writing data given from the file system to the flash memory, the adaptation layer gathers the data sectors and rearranges them suitably for the QLC flash memory.

II. Related Work

Fig. 1 shows the overall architecture of large capacity flash file system. The file system issues write commands along with logical sector numbers and data. In case of small capacity flash memories, the given logical sector number (LSN) is directly converted into a physical sector number of flash memory by the mapping algorithm provided by FTL [4]. However, in case of QLC flash memory, the following problem should be considered.

A. Problem Definition 1

In QLC flash memory, a page size is four to eight times larger than file system's data sector, although the NOP allowed within a page is only one [6]. With restricted NOP, the flash memory does not allow any additional access to a page [7]. Therefore, the QLC flash memory requires a data gathering algorithm which gathers data in the unit of a page before writing onto the flash memory.

The current basic data gathering algorithm [5] is used to gather data with same logical page number (LPN). When the file system issues a write command as "w LSN data: write data in the logical sector (LSN)", the basic data gathering algorithm calculates LPN by dividing LSN with the number of sectors per page. Each write command's data are collected in the buffer until a write command with different LPN appears.

Fig. 2 shows an example of the basic data gathering algorithm. In this figure, we assume there are four sectors within a page and one NOP per page. w 0 A, w 1 B, and w 2 C are gathered into one page, since all of them belong to LPN 0 (=0/4, =1/4, =2/4). However, when w 9 D (LPN 2 =9/4) occurs, the data in the buffer is sealed as a page and is written onto the flash memory. Finally the buffer is flushed and data D is written onto the emptied buffer. Likewise, other write commands are performed.

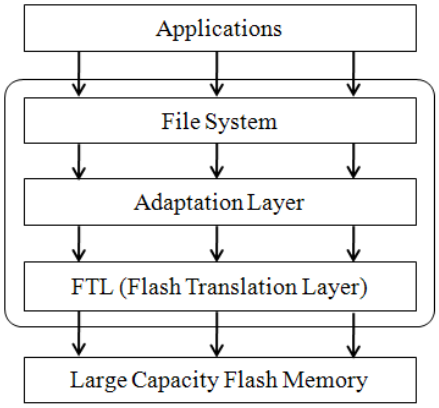


Fig. 1. Architecture of Flash File System.

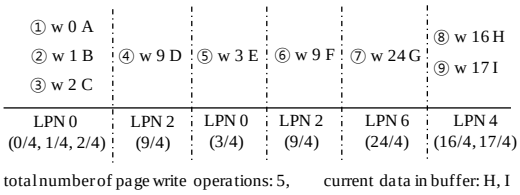


Fig. 2. Basic Data Gathering Algorithm.

The basic data gathering algorithm generates a total of five page write operations in Fig. 2, although there are only nine write commands. If each page were fully filled with data, nine write commands should generate only two to three pages since each page can store four write commands. As a solution, we propose to implement an adaptation layer between file system and FTL. The adaptation layer considers the following problem to fully gather data.

B. Problem Definition 2

The adaptation layer is required to separate random data from the gathering page. A file consists of sequential data and random data. The random data refers to file's meta data in which its LSNs are irregularly allocated and its data is frequently updated [3]. The irregular LSN allocation refers to the fact that the random data's LSNs are unlikely to be relevant to nearby LSNs. For example, in Fig. 2, the data corresponding to LSN 9 and LSN 24 are random data. Due to the update of LSN 9 and irregular allocation of LSN 24, the gathering page is sealed as a page whenever write commands with LSN 9 or LSN 24 appear. In order to solve Problem Definition 2, the adaptation layer contains an undefined buffer and two RAM pages: sequential and random.

Definition 1: The undefined buffer is an instant buffer which stores the write command that cannot be immediately decided as random or sequential. The adaptation layer requires maximum of two previous LSNs for decision; therefore, the capacity of undefined buffer is two sectors.

Definition 2: The sequential RAM page (SRP) and random RAM page (RRP) store the write commands that are defined as sequential and random respectively. The size of each RP is one page.

Each write command's LSN and its corresponding data is dynamically inserted into the undefined buffer or one of two RPs depending on the following algorithm.

1) Is the write command sequential to SRP?
2) Is the write command an update?
3) Analyze the write command with other LSNs of the undefined buffer.

First, the adaptation layer checks whether the write command belongs to SRP (<1>). If the write command does not belong to SRP, the adaptation layer checks whether the write command is an update of previous write commands. If the write command is an update, it is inserted into RRP since the update is one of characteristics of random data. However, when the write command does not belong to <1> or <2>, it is analyzed with other LSNs of the undefined buffer. The main role of <3> is to define LSNs with irregular LSN allocation as random data. The data gathering algorithm of adaptation layer is explained in detail in Section III.

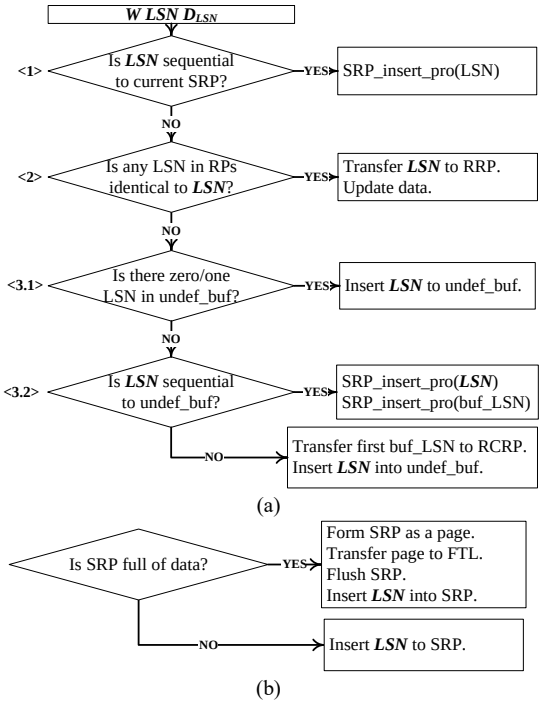


Fig. 3. (a) Data Gathering Algorithm, (b) SRP Insertion Procedure.

III. Data Gathering Algorithm of Adaptation Layer

When the system is initiated, there is no data stored in SRP or RRP. The adaptation layer requires previous LSNs to analyze the pattern of LSNs. Therefore, the adaptation layer simply accumulates the write commands with same LPN until a write command with different LPN appears. Finally the adaptation layer stores the write command with different LPN in the undefined buffer and finishes the initialization.

Example (Fig. 4(a)): For convenient understanding, we assume there are only four sectors per page. We have arbitrarily written LSN and its corresponding data within each sector as (LSN, data). Initially there is no data in SRP or RRP. (0, A), (1, B), and (2, C) are stored in SRP, since all of them belong to LPN 0 (=0/4, = 1/4, =2/4). (9, D) is stored in the undefined buffer since it belongs to LPN 2 (=9/4).

The write commands subsequent to the initialization follow the data gathering algorithm as shown in Fig. 3. Fig. 3 is a detailed view of the data gathering algorithm aforementioned in Section II. Each procedure of Section II corresponds to the procedure of Section III respectively except <3>, which is described in two parts (<3.1> and <3.2>) in Fig. 3. When the file system issues a write command, the adaptation layer checks whether the LSN is sequential to the SRP as shown in <1> of Fig. 3(a). The write command's LSN is decided as the sequential data when the differential between the write commands' LSN and the last LSN of SRP equals to the differential between two immediate last LSNs of SRP.

Example (Fig. 4(b)): w 3 E is sequential to SRP, because both result of LSN 3-LSN 2 and LSN 2-LSN 1 equals to one.

When a LSN is sequential to the SRP, the adaptation layer searches an empty sector within the SRP as explained in Fig. 3(b). If the SRP contains an empty sector, the LSN and its corresponding data can be directly inserted into the SRP. On the other hand, if the SRP is full of data, the data in SRP is sealed as a page, and it is sent to the FTL to be written onto the flash memory. Finally the SRP is flushed, and the LSN and its corresponding data are written to the SRP.

When a LSN does not belong to <1>, the adaptation layer checks whether the write command's LSN has previously appeared in the undefined buffer or RPs as explained in <2> of Fig. 3(a). If an identical LSN exists, the LSN's corresponding data is defined as the random data because one of random data's characteristics is the frequent update as mentioned in Section II. Therefore, the LSN is transferred to the RRP, and its corresponding data is updated.

Example (Fig. 4(c)): When w 9 F is issued from the file system, the adaptation layer searches for LSN 9. Due to (9, D), LSN 9 and its data F are written to the RRP and old data D is deleted.

If the write command's LSN does not belong to <1> or <2>, the adaptation analyzes the pattern of trace by comparing the write command's LSN with other undefined LSNs. Our algorithm requires two previous undefined LSNs for analysis; therefore, the LSN is temporarily stored in the undefined buffer (undef_buf) when there are less than two undefined LSNs as explained in <3.1> of Fig. 3(a).

Example (Fig. 4(d)): w 24 G does not belong to <1>, because LSN 24-LSN 3 does not equal to LSN 3-LSN 2. It does not belong to <2>, because there is no identical LSN in the undefined buffer or RPs. Thus, w 24 G must be compared to other undefined write commands. Unfortunately there is less than two LSNs in the undefined buffer so (24, G) is just stored in the undefined buffer. With same reason, next write command, w 16 H, is also stored into the undefined buffer.

When the undefined buffer contains two undefined LSNs, the adaptation layer checks whether the write command's LSN is sequential to them or not as shown in <3.2> of Fig. 3(a). The write command's LSN is considered as sequential, if the differential between the write commands' LSN and the last LSN of undefined buffer equals to the differential between two immediate last LSNs of undefined buffer. If the LSN is sequential to the undefined buffer, the undefined LSNs and write command's LSN are inserted into the SRP.

On the other hand, if the write command's LSN is not sequential to the undefined buffer, the first LSN and data of the undefined buffer is transferred to RRP, and the write command's LSN is newly inserted into the undefined buffer.

Example (Fig. 4(e)): When w 17 I is issued from the file system, the undefined buffer contains two LSNs: (24, G) and (16, H). The adaptation layer checks whether w 17 I is sequential to the undefined buffer or not. w 17 I is not sequential to the undefined buffer, because LSN 17-LSN 16 does not equal to LSN 16-LSN 24. In this case, (24, G) is transferred into RRP, and LSN 17 and its corresponding data I is newly inserted into the undefined buffer.

In Fig. 4(e), we have defined (24, G) as random data, even though LSN 24 has not appeared before. The adaptation layer defines the first LSN of the undefined buffer as random data due to the characteristic of irregular LSN allocation. The first undefined LSN is not sequential to the SRP and it does not have any chance of being a portion of sequential data in future. For example, (24, G) is not sequential to the SRP, and it is not sequential to next two commands: (16, H) and (17, I). On the other hand, second undefined LSN, LSN 16, remains in the undefined buffer, since it still has chance of being a portion of sequential data depending on the next write command (w 18 J).

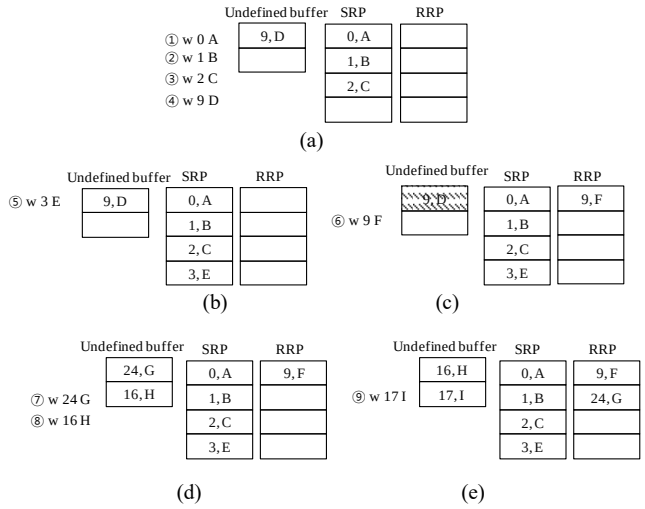


Fig. 4. (a) An Example of Initialization, (b) An Example of <1> of Fig. 3(a), (c) An Example of <2>, (d) An Example of <3.1>, (e) An Example of <3.2>.

TABLE I. NUMBER OF PAGES GENERATED BY PBM AND APRA

trace	total input	Basic data gathering algorithm (χ)	Adaptation layer (y)	difference (χ-y)
A	12,363,602	1,558,096	1,535,448	22,648
B	15,084,489	1,938,608	1,903,344	35,264
C	40,220,118	20,561,211	17,225,971	3,335,240
D	42,558,072	25,370,754	21,004,710	4,366,044
E	4,717	1,808	627	1,181
F	5,110	1,737	428	1,309
G	69,575	6,928	4,993	1,935
H	18,899	3,334	1,673	1,661

IV. Performance Evaluation

In this section, we have implemented our adaptation layer and compared it to current basic data gathering algorithm. We have analyzed the number of pages formed by each algorithm with the traces retrieved from various devices. Both algorithms are simulated on 256 Gbyte SSD, which consists of eight sectors per page and one NOP per page.

According to Table I, the adaptation layer forms smaller number of pages than current data gathering in overall environments. The adaptation layer has reduced over twenty thousand page write operations, and has reduced approximately one thousand page write operations in embedded devices. As we expected, separating the random data from the gathering page has fully filled pages with data, thus significantly reducing total number of page write operations. On the other hand, the basic data gathering algorithm formed many pages with empty sectors, because the page is likely to be sealed as a page whenever the random data interferes.

V. Conclusion

In this paper, we have dealt with the NOP restriction property of QLC flash memory. We have proposed to implement an adaptation layer between file system and FTL. It gathers and adjusts data in a unit of page so that small capacity based FTLs can be implemented on FTL without considering the NOP restriction. Furthermore, it separates random data from the gathering page, in order to reduce the number of page write operations. According to our experiment, the adaptation layer forms smaller number of pages than the current basic data gathering algorithm.

References

[1] MICRON Electronics, “Cache Programming Operations,” MICRON Electronics Technical Notes, 2022.

[2] Li-Pin Chang, “A Hybrid Approach to NAND-Flash-Based Solid-State Disks,” IEEE Transactions on Computers, 2010.

[3] Samsung Electronics, “QLC SSD,” 2022.

[4] Tatsuo Shiozawa, Hirotsugu Kajihara, Tatsuro Endo, and Kazuhiro Hiwada, “Emerging Usage and Evaluation of Low Latency FLASH,” 2020 IEEE International Memory Workshop (IMW), 2020.

[5] Mamoru Fukuchi, Shun Suzuki, Kyosuke Maeda, Chihiro Matsui, and Ken Takeuchi, “BER Evaluation System Considering Device Characteristics of TLC and QLC NAND Flash Memories in Hybrid SSDs with Real Storage Workloads,” 2021 IEEE International Symposium on Circuits and Systems (ISCAS), 2021.

[6] Yoshiki Takai, Mamoru Fukuchi, Reika Kinoshita, Chihiro Matsui, and Ken Takeuchi, “Analysis on Heterogeneous SSD Configuration with Quadruple-Level Cell (QLC) NAND Flash Memory,” 2019 IEEE 11th International Memory Workshop (IMW), 2019.

[7] R. Mativenga, J.-Y. Paik, J. Lee, T. S. Chung, and Y. Kim, “RFTL: Improving performance of selective caching-based page-level FTL through replication,” Cluster Comput., vol. 22, no. 1, pp. 1–17, 2019.

Improving Internet of Things Platform with Anomaly Detection for Environmental Sensor Data

Okyza Maherdy Prabowo[1], Suhono Harso Supangkat[2], Eueung Mulyana[3], I Gusti Bagus Baskara Nugraha[4]
STMIK AMIK Bandung, Bandung Institute of Technology, Bandung, Indonesia[1]
Smart City and Community Innovation Center, Bandung Institute of Technology, Bandung, Indonesia[2]
School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, Indonesia[3, 4]

***Abstract*—Internet of things has an essential role in various application domains. The number of Internet of Things applications makes researchers try to formulate how to design the architecture of the Internet of Things platform so that it can be used generically in various domains. Commonly used architectural designs consist of data collecting, data preprocessing, data analysis, and data visualization. However, sensor data that enters the platform often experiences anomalies such as constant values or being stuck-at zero, which are processed manually at the data preprocessing stage. In this research, we try to design an anomaly detection system on the Internet of Things platform that can automatically improve the platform's performance in detecting anomalies. In this study, we compared the False Positive Rate of several anomaly detection algorithms tested to real datasets in the environmental sensor data application domain. The results showed that the anomaly detector system on the Internet of Things platform had an optimal False Positive Rate of 0.9%.**

***Keywords*—Anomaly detection; sensor data; multivariate; Internet of Things; smart system**

I. INTRODUCTION

Recently some urban environments have extensively used internet of things (IoT) technology to perform environmental monitoring and control. The acquisition and control settings and the network protocols vary according to the urban environment's intended applications. These elements are critical to the ability of IoT networks to communicate successfully and transfer valuable data. Valuable data such as air temperature and relative humidity from inside and outdoor locations are essential for understanding the urban microclimate affecting the environmental condition. The monitoring process needs help from technology tools to automate the collection and understanding of data, for example, the internet of things platform.

The platform collects microclimate parameters from all sensor data. The platform also serves as a data management platform. The data platform architecture has a subsystem called the data analytics module. The data analytics module is responsible for analyzing the collected data. The data analytics module can be implemented through video, text, or other analytics techniques such as statistical analysis or machine learning [1]. Anomaly detection is one of the analyses performed on the data collected in an agricultural environment [2].

The common goal of anomaly detection is to find patterns in data that do not conform to "expected" or normal behavior [3]. Anomaly detection is used to monitor the environmental situation of the greenhouse [19]. When anomalous behavior is detected, an alarm can be send to the administrator to do something. Several techniques are used to detect anomalies, which can be classified into two categories: conventional and data-driven. A conventional technique like the statistical method has a long history of detecting an outlier in the data [20]. Parametric or non-parametric techniques are included in this category. The underlying distribution of the data is known for the parametric category, and the parameters are estimated using the data. Parametric methods include those based on the Gaussian distribution, the regression model, or a combination of Parametric Distributions [4]. Data-driven techniques are frequently used to refer to learning-based methods in which the lack of a robust underlying mathematical model is compensated for by the availability of large amounts of data from which useful information can be "learned." Machine learning is a large area of research with numerous application areas. Generally, it is divided into three distinct categories: supervised, unsupervised, and reinforcement learning. Additionally, due to technological advancements, deep learning is gaining popularity. Numerous machine learning techniques are frequently given a deep learning orientation or are combined with deep learning [18].

Several algorithms for detecting sensor abnormalities are used in agricultural environments. Several neural network algorithms were used, including artificial neural networks, autoencoders, recurrent neural networks, and long short-term memory. However, in complex environments where a clear variation pattern for some greenhouse parameters is complicated, environmental anomalies are rarely captured or recognized by univariate sensor data analysis or single machine learning models [5]. A multivariate anomaly detection approach is needed to be explored in Internet of Things area [16]. The anomaly detector system proposed in this study uses a GRU-based Variational Autoencoder. Guo proposed this method to handle IoT sensor data in Smart City [7]. The advantage of the GRU-based anomaly detection system is its reliability in discovering the data correlation and dependencies [14]. There are still weaknesses in actual labelling, which will be improved in this study by involving human knowledge as part of a multivariate anomaly detection system. The main contribution in this paper are summarized as follows:

- Anomaly detector framework for the Internet of Things platform handling type of sensors error combining data-driven and knowledge-driven.
- Algorithm comparison results that fit the Internet of things platform performing real-world datasets in the greenhouse system in the tropical country.

The rest of this article is organized as follows. The Gated Recurrent Unit (GRU) and Variational Autoencoder (VAE)-based anomaly detector are described in Section II, along with the proposed architecture combining GRU-VAE and Human-in-the-loop method. By incorporating knowledge into data-driven techniques, we can increase the detection rate of interesting anomalies [11]. Section III contains the results and discussion for three datasets. Finally, Section IV contains some concluding remarks.

II. System Model

In this section, the Gated Recurrent Unit and Variational Autoencoder-based anomaly detection are given along with the proposed model.

A. Gated Recurrent Unit

The GRU, or gated recurrent unit, is an improvement over the RNN, or recurrent neural network. It was first introduced by Cho et al. in 2014[6]. GRUs are strikingly similar to Long-Short-Term Memory (LSTM). Similar to LSTM, GRU utilizes gates to control the information flow. In comparison to LSTM, they are a more recent development. Consequently, they outperform LSTM and have a more straightforward architecture, as referenced in Fig. 1.

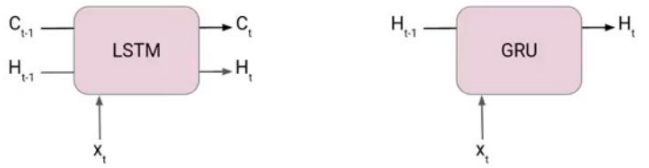


Fig. 1. Illustration Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) Cell.

At each timestamp t, it accepts the input X_t and the hidden state H_{t-1} from the previous timestamp t-1. The new hidden state H_t is then output and passed to the subsequent timestamp. As shown in Figures 1 and 2, a GRU cell consists primarily of two gates instead of three in an LSTM cell. The reset gate comes first, followed by the update gate. The reset gate is responsible for the network's short-term memory, known as its hidden state (H_t). The equation for the reset gate is as follows.

$$r_t = \sigma\ (x_t * U_r + H_{t-1} * W_r) \quad (1)$$

The update gate for long-term memory is illustrated below, along with the gate's equation.

$$u_t = \sigma\ (x_t * U_u + H_{t-1} * W_u) \quad (2)$$

The only distinction is between the weight metrics U_u and W_u. To locate the hidden state H_t in GRU, a two-step procedure is used. The first step is to create a candidate's hidden state. The hidden gate formula is described.

$$H_t = tanh\ (x_t * U_g + (r_t \circ H_{t-1}) * W_u) \quad (3)$$

It multiplies the input and hidden state from the previous timestamp t-1 by the output of the reset gate r_t. This information is then passed to the tanh function, which returns the hidden state of the candidate. Important in this equation is how we use the reset gate's value to limit the previous hidden state's influence on the candidate state. If r_t equals 1, the previous hidden state H_{t-1} is evaluated in its entirety. Similarly, if r_t is 0, the information from the previous hidden state is completely disregarded. After determining the candidate state, it is used to generate the current hidden state Ht. The Update gate enters the fray at this point. This equation is highly intriguing because, unlike LSTM, we control the historical information in H_{t-1} and the new information in the candidate state with a single update gate.

$$H_t = u_t \circ H_{t-1} + (1 - u_t) * H_t) \quad (4)$$

Now, if U_t is close to 0, the first term in the equation will vanish, implying that the new hidden state will contain little information about the previous hidden state. On the other hand, the second part becomes nearly identical to the first, which implies that the hidden state at the current timestamp will contain only information from the candidate state. Guo uses GRU cells in both the encoder and the decoder to discover the data correlation and dependency [7].

B. Variational Autoencoder based Anomaly Detection

Anomaly detection is one of those domains where machine learning has had such a profound impact that it is almost axiomatic that anomaly detection systems must be based on some type of automatic pattern learning algorithm as opposed to a set of rules or descriptive statistics (though many reliable anomaly detection systems operate using such methods very successfully and efficiently). Combining Bayesian inference with an AE framework, VAE is a probabilistic model. As opposed to a reconstruction error, a VAE-based anomaly detection model generates a probabilistic measure for the anomaly score [14]. Reconstruction probabilities are more objective and principled than reconstruction errors because they do not require modeling specific thresholds for judging anomalies. In particular, VAE assumes that a large number of complex data distributions can be described by a smaller set of latent variables with more straightforward probability density distributions. Thus, the objective of VAE is to find a low-dimensional representation of the latent variables in the input data.

The VAE is distinct from conventional autoencoders because it is probabilistic and generative. The VAE generates partially random outputs (even after training) and can also generate new data similar to the data on which it was trained. The VAE is structurally similar to a conventional autoencoder at a high level. However, the encoder acquires additional coding; specifically, the VAE acquires mean and standard deviation coding. The VAE then generates the latent variables, z, by randomly sampling from a Gaussian distribution with the same mean and standard deviation as the encoder. To reconstruct the input, these latent variables are "decoded." The architecture is visualized by Fig. 2.

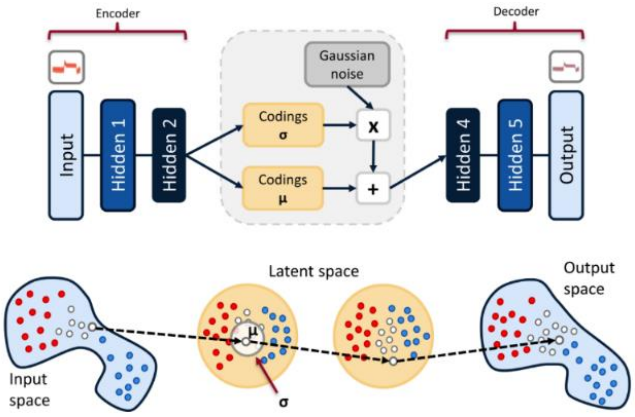


Fig. 2. A Variational Autoencoder Architecture (Top), and an Example of a Data Sample Going through the VAE (Bottom).

C. Human Knowledge-Driven

Human-driven machine intelligence is also known as "human in the loop." Human-in-the-loop (HITL) is a subfield of artificial intelligence that combines human and machine intelligence to create machine learning models [12]. In a conventional human-in-the-loop approach mentioned in Fig. 3, people are involved in a virtuous circle in which they train, tune, and test a specific algorithm. Humans initially assign labels to data, which provides a model with high-quality (and massive) training data. A machine learning algorithm learns to make decisions based on this data. Afterward, humans fine-tune the model. Humans typically score data to account for overfitting, to teach a classifier about edge cases, or to add new categories to the model's scope. Individuals can evaluate and validate a model's outputs, especially when an algorithm is uncertain about a judgment or overconfident about an incorrect decision [10].

D. Proposed Architecture

The proposed architecture for anomaly detection combines data-driven methods, specifically a GRU-based Variational Autoencoder, with expert-provided knowledge, as referenced in Fig. 4. The GRU-based variational autoencoder performs the anomaly detection process on the provided dataset and then compares it to the expert's knowledge [13]. On the greenhouse dataset, this knowledge-driven approach is used. The following diagram illustrates a multivariate anomaly detector's general architecture. The GRU-based VAE algorithm detects anomalies from multivariate sensors. Similarly, experts interpret anomalies in multivariate data.

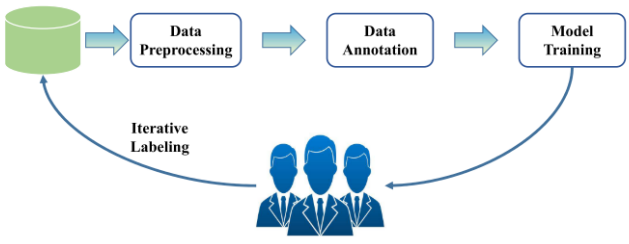


Fig. 3. The Development Cycle of Model [10].

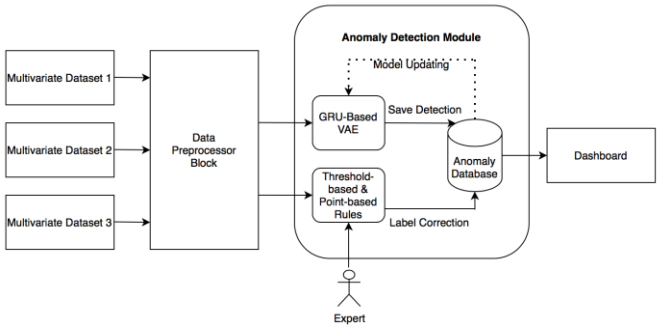


Fig. 4. Fusion Architecture Combining Data-driven and Knowledge- Driven.

This multivariate dataset is derived from three datasets: Intel Berkeley Dataset, indoor greenhouse sensors, and outdoor greenhouse sensors. Greenhouse sensors are being installed in Bandung, Indonesia, in the tropics. The data is cleaned of empty values in the preprocessing subsystem, and each dataset is resampled every 20 minutes. Following that, the feature scaling process is carried out, which is rescaling features to make them more suitable for training [15].

Data Preprocessing blocks are used to carry out several data management processes such as joining data, removing missing values, separating data according to sensor categories, resampling time and then entering into two different system blocks, namely data-driven block and human knowledge-driven block[17].

A data-driven anomaly detection architecture will be proposed in this study, which will make use of a GRU-based variational autoencoder with multivariate time-series data as input. First, greenhouse data is loaded from the database and split into two dataset categories, indoor module, and outdoor module. Each sensor contains four sensor variables, a battery sensor, a humidity sensor, and two temperature sensors.

The GRU input accepts four sensor inputs, each of which is connected to 150 cells in the first layer. Then, the output of the second layer is connected to the second layer's 100 inputs.

The human-in-the-loop method is used for knowledge-driven anomaly detection. Experts make label recommendations based on data using threshold-based and point-based methods. This threshold-based approach will be used to improve the anomaly detection threshold generated by the GRU-based VAE in the future. At the same time, the point-based is used by iteratively providing data to the expert and then labeling the points. The point-based anomaly detection process is measured by the amount of time it takes the expert to label it versus the amount of anomaly it takes to help the expert label it.

III. Experiment Result and Discussion

In this section, the proposed GRU-based VAE model is evaluated using the Intel Berkeley, greenhouse indoor sensor, and greenhouse outdoor sensor datasets. Shamshiri proposed microclimate parameters to be evaluated [8]. Four criteria, namely accuracy, the area under curve (AUC), true positive rate, and false positive rate, are used to evaluate the performance. All experiments were run on the Google Colab

with Intel Xeon @ 2.2GHz, 12 GB Ram, and 12GB NVIDIA Tesla K80 GPU. The algorithm was implemented using Python in Keras and Scikit-learn. The expert knowledge-driven method is evaluated using greenhouse indoor sensor and greenhouse outdoor sensor dataset. The expert will be given a set of preprocessed data and labeled them. The time execution and the number of anomaly data will be compared as the evaluation metric. The expert knowledge-driven method performs better when the time required to determine the number of known anomalies is reduced.

A. Intel Berkeley Dataset

This dataset was compiled using data from 54 sensors installed in the Intel Berkeley Research lab between February 28th and April 5th, 2004 [9]. Every 31 seconds, it contains time stamped topology information, humidity, temperature, light, and voltage values. For various sensors, there are some missing values at certain timestamps. To begin, we use the linear interpretation method to fill in the gaps. Then, every 20 minutes, we sample it and use the average as input. To balance the type of sensor that fits the greenhouse sensor, we only take temperature, humidity, light, and voltage sensors. In the meantime, we normalize the data.

With GRU-based VAE, the average testing MAE is 0.04 and MSE is 0.01 with training time 194s. Table I shows GRU-Based VAE performance. The testing accuracy, Area Under Curve (AUC), Optimal True Positive Rate (Opt. TPR) and False Positive Rate (Opt. FPR) of 81%, 69%, 4%, and 1.8% for light sensor, 39%, 89%, 14.9% and 0% for voltage sensor, 34%, 84%, 5% and 1.6% for humidity sensor and 25%, 82%, 8.4% and 1.4% for temperature sensor, respectively.

TABLE I. GRU-BASED VAE PERFORMANCE

GRU-Based VAE				
Sensor Type	**Accuracy**	**AUC**	**Opt. TPR**	**Opt. FPR**
Light Sensor	81%	69%	4%	1.8%
Voltage Sensor	39%	89%	14.9%	0%
Humidity Sensor	34%	84%	5%	1.6%
Temperature Sensor	25%	82%	8.4%	1.4%

With Gaussian Mixture Model, the testing accuracy, Area Under Curve, Optimal True Positive Rate (Opt. TPR) and False Positive Rate (Opt. FPR) of 61%, 65%, 3.7%, and 1.2% for light sensor, 27%, 94%, 12.8% and 0% for voltage sensor, 32%, 84%, 4.9%, and 1.7% for humidity sensor and 23%, 49%, 0% and 23.7% for temperature sensor, respectively. Table II summarizes the evaluation result.

TABLE II. GAUSSIAN MIXTURE MODEL PERFORMANCE

Gaussian Mixture Model				
Sensor Type	**Accuracy**	**AUC**	**Opt. TPR**	**Opt. FPR**
Light Sensor	70%	65%	3%	1.8%
Voltage Sensor	73%	7%	0%	1%
Humidity Sensor	68%	16%	2%	4.9%
Temperature Sensor	25%	51%	2.4%	0%

With K-Means, the testing accuracy, Area Under Curve, Optimal True Positive Rate (Opt. TPR) and False Positive Rate (Opt. FPR) of 70%, 65%, 3.3%, and 1.8% for light sensor, 73%, 7%, 0% and 1.29% for voltage sensor, 68%, 16%, 1.7%, and 4.9% for humidity sensor and 78%, 51%, 2.4% and 0% for temperature sensor, respectively. Table III summarizes the evaluation result.

TABLE III. K-MEANS PERFORMANCE

K-Means				
Sensor Type	**Accuracy**	**AUC**	**Opt. TPR**	**Opt. FPR**
Light Sensor	70%	65%	3%	1.8%
Voltage Sensor	73%	7%	0%	1.3%
Humidity Sensor	68%	16%	1.7%	4.9%
Temperature Sensor	78%	51%	2.4%	0%

Based on the multivariate correlation between all sensors, there are no concurrent anomalies on the four sensors. There are 0.625% concurrent anomalies on the three sensors. There are 8.125% concurrent anomalies on the two sensors.

B. Greenhouse Indoor Dataset

This dataset was compiled using data from four sensors installed outside the Greenhouse Smart City Living Lab between October 16th, 2020, and July 19th, 2021. It contains timestamped information about the topology every 60 seconds, as well as humidity, two temperatures with distinct locations, and voltage values. There are some values missing at certain timestamps for various sensors. To begin, we will fill in the gaps using the linear interpretation technique. After that, it is sampled every 20 minutes, and the average is used as an input. We take two temperatures with a different locations, humidity, and voltage sensors to balance the sensor type that fits the greenhouse sensor. Meanwhile, we standardize the data.

With GRU-based VAE, the average testing MAE is 0.015 and MSE is 0.16 with training time 196s. Table IV shows GRU-Based VAE performance. The testing accuracy, Area Under Curve (AUC), Optimal True Positive Rate (Opt. TPR) and False Positive Rate (Opt. FPR) of 82%, 58%, 7%, and 3.2% for battery sensor, 93%, 74%, 8% and 1.4% for humidity sensor, 70%, 68%, 2.1% and 0.9% for temperature sensor DS-type and 71%, 70%, 6.4% and 2.8% for temperature sensor SHT-type, respectively.

TABLE IV. GRU-BASED VAE PERFORMANCE

GRU-Based VAE				
Sensor Type	**Accuracy**	**AUC**	**Opt.TPR**	**Opt. FPR**
Battery Sensor	82%	58%	7%	3.2%
Humidity Sensor	93%	74%	8%	1%
Temperature Sensor DS	70%	68%	2.1%	0.9%
Temperature Sensor SHT	71%	70%	6.4%	2.8%

With Gaussian Mixture Model, the testing accuracy, Area Under Curve, Optimal True Positive Rate (Opt. TPR) and False Positive Rate (Opt. FPR) of 64%, 55%, 5.5%, and 2.9% for battery sensor, 98%, NaN, 0% and 1.8% for humidity sensor,

66%, 69%, 0.2%, and 3.4% for temperature sensor DS-type and 69%, 72%, 1.4% and 8.8% for temperature sensor SHT-type, respectively. Table V summarizes the evaluation result.

TABLE V. GAUSSIAN MIXTURE MODEL PERFORMANCE

Gaussian Mixture Model				
Sensor Type	**Accuracy**	**AUC**	**Opt. TPR**	**Opt. FPR**
Battery Sensor	64%	55%	5.5%	2.9%
Humidity Sensor	98%	NaN	0%	2%
Temperature Sensor DS	66%	69%	0.2%	3.4%
Temperature Sensor SHT	69%	72%	1.4%	8.8%

With K-Means, the testing accuracy, Area Under Curve, Optimal True Positive Rate (Opt. TPR) and False Positive Rate (Opt. FPR) of 51%, 54%, 4.9%, and 2.7% for battery sensor, 98%, 100%, 100% and 1.8% for humidity sensor, 68%, 69%, 3.3%, and 2% for temperature sensor DS-type and 70%, 71%, 8.7% and 1.6% for temperature sensor SHT-type, respectively. Table VI summarizes the evaluation result.

TABLE VI. K-MEANS PERFORMANCE

K-Means				
Sensor Type	**Accuracy**	**AUC**	**Opt. TPR**	**Opt. FPR**
Battery Sensor	51%	54%	4.9%	2.7%
Humidity Sensor	98%	100%	100%	1.8%
Temperature Sensor DS	68%	69%	3.3%	2%
Temperature Sensor SHT	70%	71%	8.7%	1.6%

Based on the multivariate correlation between all sensors, there are 0.024% concurrent anomalies on the four sensors. There are 0.16% concurrent anomalies on the three sensors. There are 1.85% concurrent anomalies on the two sensors.

C. Greenhouse Outdoor Dataset

This dataset was compiled using data from 4 sensors installed in the Greenhouse Smart City Living Lab between October 16th, 2020, and July 19th, 2021. It also contains timestamped information about the topology every 60 seconds, as well as humidity, two temperatures with distinct locations, and voltage values. There are some values missing at certain timestamps for various sensors. To begin, we will fill in the gaps using the linear interpretation technique. After that, it is sampled every 20 minutes, and the average is used as an input. We take two temperatures with a different locations, humidity, and voltage sensors to balance the sensor type that fits the greenhouse sensor. Meanwhile, we standardize the data.

With GRU-based VAE, the average testing MAE is 0.05 and MSE is 0.007 with training time 72s. Table I shows GRU-Based VAE performance. The testing accuracy, Area Under Curve (AUC), Optimal True Positive Rate (Opt. TPR) and False Positive Rate (Opt. FPR) of 86%, 59%, 5.3%, and 1.1% for battery sensor, 78%, 60%, 1.5% and 1.5% for humidity sensor, 68%, 69%, 4.7% and 1% for temperature sensor DS-type and 69%, 63%, 1.2% and 0.6% for temperature sensor SHT-type, respectively.

TABLE VII. GRU-BASED VAE PERFORMANCE

GRU-Based VAE				
Sensor Type	**Accuracy**	**AUC**	**Opt. TPR**	**Opt. FPR**
Battery Sensor	86%	59%	5.3%	1.1%
Humidity Sensor	78%	60%	1.5%	1.5%
Temperature Sensor DS	68%	69%	4.7%	1%
Temperature Sensor SHT	69%	63%	1.2%	0.6%

With Gaussian Mixture Model, the testing accuracy, Area Under Curve, Optimal True Positive Rate (Opt. TPR) and False Positive Rate (Opt. FPR) of 61%, 57%, 3.3%, and 0.7% for battery sensor, 29%, 65%, 1.7% and 0.9% for humidity sensor, 68%, 67%, 5.4%, and 0.6% for temperature sensor DS-type and 73%, 62%, 1% and 0.7% for temperature sensor SHT-type, respectively. Table VIII summarizes the evaluation result.

TABLE VIII. GAUSSIAN MIXTURE MODEL PERFORMANCE

Gaussian Mixture Model				
Sensor Type	**Accuracy**	**AUC**	**Opt. TPR**	**Opt. FPR**
Battery Sensor	61%	57%	3.3%	0.7%
Humidity Sensor	29%	65%	1.7%	0.9%
Temperature Sensor DS	68%	67%	5.4%	0.6%
Temperature Sensor SHT	73%	62%	1%	0.7%

With K-Means, the testing accuracy, Area Under Curve, Optimal True Positive Rate (Opt. TPR) and False Positive Rate (Opt. FPR) of 48%, 57%, 0.7%, and 2.6% for battery sensor, 73%, 62%, 1.9% and 1.3% for humidity sensor, 68%, 67%, 5.5%, and 0.6% for temperature sensor DS-type and 72%, 63%, 0.7% and 0.9% for temperature sensor SHT-type, respectively. Table IX summarizes the evaluation result.

TABLE IX. K-MEANS PERFORMANCE

K-Means				
Sensor Type	**Accuracy**	**AUC**	**Opt. TPR**	**Opt. FPR**
Battery Sensor	48%	57%	0.7%	2.6%
Humidity Sensor	73%	62%	1.9%	1.3%
Temperature Sensor DS	68%	67%	5.5%	0.6%
Temperature Sensor SHT	72%	63%	0.7%	0.9%

Based on the multivariate correlation between all sensors, there are 0.13% concurrent anomalies on the four sensors. There are 0.19% concurrent anomalies on the three sensors. There are 0.58% concurrent anomalies on the two sensors.

D. Threshold-based & Point-based Human Knowledge Driven

This paper introduces point-based anomaly detection as part of a proposed method for determining how human-in-the-loop evaluation can be performed. This proposed method describes how an expert can provide anomaly recommendations through the threshold and point annotations. An agricultural expert was involved in determining the point anomaly in the Greenhouse Smart City Living Lab context in this greenhouse case study.

Experts are given raw data in stages, and the amount of raw data given is used to determine how much raw data is required to make it easier for experts to annotate anomalies. The time the data is displayed before the expert can provide annotations is then calculated. Based on the results of tests with data ranging from n = 1 to n = 100, it was discovered that the optimal expert produced the fastest results with n = 5 and an annotation time of 7.12 seconds. That is, the expert requires a minimum of five data samples in order to draw an annotation conclusion. Human knowledge is used as an adaptive threshold in the threshold-based approach, which can replace sigma, which is currently used as a threshold limit. With human knowledge stored in the database, the anomaly detection process will become more adaptive by adjusting the context or rules provided by humans based on point annotations or specific conditions such as a crop disease [20].

E. Discussion

GRU-based VAE has performed well in detecting anomalies, particularly the relationship between the detected variables. However, GRU-based VAE does not produce the best results in some datasets because it necessitates layer adjustments based on the data conditions. However, the deficiency in the anomaly detection process is compensated for by the assistance of human knowledge. Unlike the other algorithms, it has not been able to demonstrate the correlation between anomalies from multiple sensors simultaneously. However, a more detailed assessment of this correlation is required. Correlation is only indicated in this study by the classifications of no correlation or correlates.

This study also proposes a new metric for measuring human-in-the-loop by comparing the amount of data required for annotation and the time it takes the expert to annotate. The comparison curve of the anomaly n and the required time t is generally close to a quadratic function, making it difficult for the expert to annotate due to a lack of data. However, having too much data will also make it difficult for the expert.

IV. CONCLUSION

The greenhouse, outfitted with sensors, generates a large amount of data that must be processed. Of course, the data cannot be separated from anomalies, which may be an anomaly that must be removed because it corrupts the data, or the anomaly may represent hidden information that can be used to make future decisions. The Gated Recurrent Unit-based Variational Autoencoder is proposed in this study as an anomaly detection algorithm capable of detecting anomalies in the multivariate term. This algorithm is also a component of the anomaly detection architecture, which is enhanced by threshold-based and point-based anomaly detection based on human knowledge, which can improve anomaly detection performance.

This anomaly detection model and architecture were evaluated using the Intel Berkeley Lab Dataset, Greenhouse Smart City Living Lab Dataset, and indoor and outdoor sensors. The evaluation results demonstrate that the proposed model is superior at detecting multivariate anomalies and identifying variable correlation. Our proposed architecture using GRU-based VAE and expert feedback can examine correlations between multivariable time series data. The human knowledge module enhances the performance of the GRU-based VAE by correcting false alarms and detecting errors.

Future works are necessary to validate the kinds of conclusions that can be drawn from this research. For example, it is necessary to measure point-based and threshold data for future case studies that may generate different curves. In addition, it is anticipated that this human knowledge will be utilized automatically to enhance the GRU-based VAE anomaly detection model in the future research.

REFERENCES

[1] Erhan, L., Ndubuaku, M., Di Mauro, M., Song, W., Chen, M., Fortino, G., Bagdasar, O., & Liotta, A. (2021). Smart anomaly detection in sensor systems: A multi-perspective review. Information Fusion, 67, 64–79. https://doi.org/10.1016/j.inffus.2020.10.001.

[2] Ou, C. H., Chen, Y. A., Huang, T. W., & Huang, N. F. (2020). Design and Implementation of Anomaly Condition Detection in Agricultural IoT Platform System. International Conference on Information Networking, 2020-Januari, 184–189. https://doi.org/10.1109/ICOIN48656.2020.9016618.

[3] Liu, Y., Pang, Z., Karlsson, M., & Gong, S. (2020). Anomaly detection based on machine learning in IoT-based vertical plant wall for indoor climate control. Building and Environment, 183. https://doi.org/10.1016/j.buildenv.2020.107212.

[4] Farzad, A., & Gulliver, T. A. (2020). Unsupervised log message anomaly detection. ICT Express, 6(3), 229–237. https://doi.org/10.1016/j.icte.2020.06.003.

[5] Liu, Y., Pang, Z., Karlsson, M., & Gong, S. (2020). Anomaly detection based on machine learning in IoT-based vertical plant wall for indoor climate control. Building and Environment, 183, 107212. https://linkinghub.elsevier.com/retrieve/pii/S0360132320305837.

[6] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. 1–9. http://arxiv.org/abs/1412.3555.

[7] Guo, Y., Ji, T., Wang, Q., Yu, L., Min, G., & Li, P. (2020). Unsupervised Anomaly Detection in IoT Systems for Smart Cities. IEEE Transactions on Network Science and Engineering, 7(4), 2231–2242. https://doi.org/10.1109/TNSE.2020.302754.

[8] Shamshiri, R. R., Bojic, I., van Henten, E., Balasundram, S. K., Dworak, V., Sultan, M., & Weltzien, C. (2020). Model-based evaluation of greenhouse microclimate using IoT-Sensor data fusion for energy efficient crop production. *Journal of Cleaner Production*, *263*, 121303. https://doi.org/10.1016/j.jclepro.2020.121303.

[9] Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings, Ml, 1–14.

[10] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2021). A Survey of Human-in-the-loop for Machine Learning. http://arxiv.org/abs/2108.00941.

[11] Van Der Stappen, A., & Funk, M. (2021). Towards Guidelines for Designing Human-in-the-Loop Machine Training Interfaces. International Conference on Intelligent User Interfaces, Proceedings IUI, 514–519. https://doi.org/10.1145/3397481.3450668.

[12] McBride, N. (2021). Human in the loop. Journal of Information Technology, 36(1), 77–80. https://doi.org/10.1177/0268396220946055.

[13] Steenwinckel, B., De Paepe, D., Vanden Hautte, S., Heyvaert, P., Bentefrit, M., Moens, P., Dimou, A., Van Den Bossche, B., De Turck, F., Van Hoecke, S., & Ongenae, F. (2021). FLAGS: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge with machine learning. Future Generation Computer Systems, 116, 30–48. https://doi.org/10.1016/j.future.2020.10.015.

[14] Guo, Y., Liao, W., Wang, Q., Yu, L., Ji, T., & Li, P. (2018). Multidimensional Time Series Anomaly Detection: A GRU-based Gaussian Mixture Variational Autoencoder Approach. Proceedings of

Machine Learning Research, 95(2001), 97–112. http://proceedings.mlr.press/v95/guo18a.html.

[15] Vilenski, E., Bak, P., & Rosenblatt, J. D. (2019). Multivariate anomaly detection for ensuring data quality of dendrometer sensor networks. Computers and Electronics in Agriculture, 162(November 2018), 412–421. https://doi.org/10.1016/j.compag.2019.04.018.

[16] Cook, A. A., Misirli, G., & Fan, Z. (2020). Anomaly Detection for IoT Time-Series Data: A Survey. IEEE Internet of Things Journal, 7(7), 6481–6494. https://doi.org/10.1109/JIOT.2019.2958185.

[17] Eredics, P., & Dobrowiecki, T. P. (2011). Data cleaning for an intelligent greenhouse. SACI 2011 - 6th IEEE International Symposium on Applied Computational Intelligence and Informatics, Proceedings, 293–297. https://doi.org/10.1109/SACI.2011.5873017.

[18] Mehra, M., Saxena, S., Sankaranarayanan, S., Tom, R. J., & Veeramanikandan, M. (2018). IoT based hydroponics system using Deep Neural Networks. Computers and Electronics in Agriculture, 155(October), 473–486. https://doi.org/10.1016/j.compag.2018.10.015.

[19] Castañeda-Miranda, A., & Castaño-Meneses, V. M. (2020). Internet of things for smart farming and frost intelligent control in greenhouses. Computers and Electronics in Agriculture, 176(May), 105614. https://doi.org/10.1016/j.compag.2020.105614.

[20] Skelsey, P. (2021). Forecasting Risk of Crop Disease with Anomaly Detection Algorithms. Phytopathology®, PHYTO-05-20-018. https://doi.org/10.1094/phyto-05-20-0185-r.

Math Balance Aids based on Internet of Things for Arithmetic Operational Learning

Novian Anggis Suwastika[1]*, Yovan Julio Adam[2], Rizka Reza Pahlevi[3], Maslin Masrom[4]
Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia[1, 4]
School of Computing, Telkom University, Bandung, Indonesia[1, 2, 3]

***Abstract*—Industry 4.0 has changed various aspects of human life towards an era heavily influenced by information technology. The impact of industry 4.0 on the education sector has led to the emergence of the term education 4.0. The Internet of Things (IoT) is one of the pillars of industry 4.0. With its capabilities, IoT can provide opportunities to develop innovations in the field of education. Several studies show that teaching aids can improve the quality of learning and learning outcomes. In Indonesia, mathematics is a compulsory subject taught from elementary school to higher education. Previous studies that used mathematical (math) balance aids to help students learn mathematical operations showed positive correlations between the learning process and student learning outcomes in the materials related to arithmetic operations. This study aims to develop an IoT-based mathematical balance tool to support three education 4.0 trends: remote access, personalization, and practice and feedback. This study used modifications of Fahmideh and Zogwhi's IoT development method. There are five phases of IoT development: initialization phase, analysis phase, design phase, implementation phase, and evaluation phase. From each phase of IoT development, IoT-based mathematical balance assistance systems have been successfully built and it complies with the functionality described in the analysis phase. The system performance also shows optimal results with 100% accuracy for reading the student's learning activities. Moreover, it uses less than 10 seconds for processing 1000 data requests.**

Keywords*—Arithmetic operation; education 4.0; internet of things development; math balance aids

I. Introduction

In 2009, Jerald summarized that automation, globalization, workplace change, and policies increasing personal responsibility had changed the trend of school student skills known as 21st-century skills [1]. In 2011, the High-Tech Strategy project organized by the German government brought up the term "Industry 4.0" or IR4.0 to promote a new manufacturing revolution based on computerization and the potential of new technologies [2]. The impact of IR4.0 on the education sector has resulted in the emergence of the term "Education 4.0" as a response to the need for educational outcomes that are aligned and meet the needs of IR4.0 [3], [4]. Fisk defines education 4.0 in nine trends, namely (1) anywhere anytime (diverse time and place), (2) personalized learning, (3) free choice (flexible delivery), (4) project-based (modular and projects), (5) field experiences (practical application), (6) data interpretation (why/where not what/how), (7) completely change of exams (evaluated not examined), (8) student ownership, (9) mentoring and peers [3]. Miranda formulates the core components of education 4.0, which are composed of (1) competencies, (2) learning methods, (3) information and communication technologies (ICT), and infrastructure [4].

ICT has a crucial role in the educational system in providing access, distribution, calculation, and collection of information [5]–[7]. ICT revolutionizes the traditional paradigm to a student-centered model, changes the way of teaching, collaborates between educational stakeholders, and creates various innovations in teaching and learning activities. Miranda classifies ICT into two groups based on the basic components and the combination of the basic components of ICT that results in various innovations and new services. One of the technologies from ICT in Miranda's core education 4.0 core component is the Internet of Things (IoT). The ability of IoT to connect various "things" through the Internet network and the ability of IoT to be programmed with computational intelligence provide opportunities to develop innovations in the education sector[8]–[10].

In the field of education, Kassab et al. have summarized in their publication a systematic literature review on the benefits and challenges of IoT in education. [11]. IoT, with its capabilities, can provide various innovation opportunities to improve the quality of teaching and learning activities at various levels of education. One of the criteria used by Kassab in conducting a systematic literature review is Ambrose's seven effective learning principles, namely previous knowledge, knowledge organization, course climate, motivation, mastery, practice and feedback, and self-directed learning [12]. In another study, Saeed et al. summarized the advantages of implementing IoT in higher education to build a smart campus (smart parking, smart inventory, smart student tracking, etc.), as the main component in smart classrooms (interactive whiteboards, attendance tracking system, wireless door locks, etc.), and develop intelligent labs (integrating IoT with LMS) [13]. Although in many publications, the focus of education 4.0 is explicitly for the secondary school or higher education level, in its implementation, the concept and technology of education 4.0 can be applied at various levels of education, starting from the elementary school level.

In Indonesia, mathematics is one of the compulsory subjects taught to students from the elementary school level. Mathematics is crucial for building children's ability to think logically, problem-solving, creativity, and cultural development [14]. However, many studies have shown that mathematics in primary schools in Indonesia is a difficult and frightening subject. This condition occurs due to various factors such as general factors (physiological, pedagogic,

*Corresponding Author.

intellectual, infrastructure, school environment) or particular factors (difficulty understanding concepts, deficiency of arithmetic operation skills, or difficulty understanding context) [15]. Teaching aids in mathematics subjects help motivate students to learn, provide a concrete picture of the concepts of mathematics lessons that tend to be abstract, and help provide an overview of the relationship between mathematics and the surrounding natural conditions [16]. Examples of mathematics teaching aids are mathematical balance aids to help students understand the material for basic operations in mathematics [17]–[22].

Integrating math balance teaching aids with IoT technology will provide added value, especially to support the education 4.0 trend and support one of Ambrose's effective learning principles, namely practice, and feedback. The integration process between teaching aids and IoT has challenges in meeting the characteristics of education 4.0. These challenges include how to identify system functionality (such as remote access, personalized learning, practice, feedback, etc.), how to determine variables and how to measure performance, how to select system components, how to system architecture, how to implement the system, and how to evaluate the system based on system objectives. This study aims to design and implement mathematics teaching aids using IoT technology that can support the education 4.0 trend: remote access, personalized learning, practice, and feedback. This research provides guidelines for developing an IoT-based teaching aid system in various types of subjects or courses to support the implementation of education 4.0.

The structure of this article consists of five sections. The first section discusses the world of today's education and the development of IoT technology, problems in the world of education, especially for learning mathematics in elementary schools in Indonesia, as well as a statement about the purpose of this research. The second section describes the research method applied in this research, starting from the initiation phase to the evaluation phase. The results of system implementation are discussed in the implementation section. The evaluation and discussion section explains the evaluation of the implementation results and discusses the achievements and opportunities for system development in the future. The last section is a conclusion that summarizes the entire section of this paper.

II. Literature Review

This section summarizes the latest opportunities and conditions for the development of IoT in the field of education. There have been several publications that discuss the implementation of IoT with teaching aids at various levels of education. The integration of IoT technology with game tools that support children's gross motoric training for children aged 4-6 years was published by Shonia et al. which integrates bag toss games with IoT [23]. Shonia et al. implement the system and test the system from the aspect of functionality and system performance. System functionality is tested by checking whether the devices can work according to their functions. Testing system functionality includes testing the microcontroller (whether the microcontroller can run a program to calculate game scores), infrared sensor testing (whether the infrared sensor can detect beanbags), testing the Wi-Fi module (whether the Wi-Fi module can send results to the IoT platform via the internet), and LED testing (whether the LED can provide color according to the bag toss hole). Meanwhile, the system performance was tested based on data communication criteria and the accuracy of reading student activity data. System testing involved four children in measuring the condition of their motor development.

Another research study on IoT to support education was published by Rahmanto et al. and by Jati et al., who use the IoT-integrated variant of the hopscotch game [24], [25]. Rahmanto et al. integrate IoT with hopscotch built using puzzle carpet. The IoT component consists of a microcontroller (using Arduino Mega 2560), a vibration detection sensor (piezoelectric ceramic vibrate sensor), a child activity indicator (buzzer), and a communication module (ESP8255-01). Meanwhile, Jati et al. built hopscotch using aluminum foil and capacitive sensors to detect children's activities. The system architecture of the two studies is the same. However, in the research of Jati et al., the game leveling system has been implemented. Game leveling can be managed using a website application. The two systems built were tested based on system functionality and system performance. Both publications test the functionality of the system by testing each hardware component of the system. While the performance test, Rahmanto et al. tested the aspects of reading speed and accuracy of the assessment. Meanwhile, Jati et al., apart from testing system delay and system accuracy, also tested system gameplay. These two studies did not measure children's motor development achievement.

Wajdi et al implemented IoT in drop-box games to help children's gross motor development [26]. The drop box game utilizes a 32cm x 40cm board with 3-5 holes. The ball is placed on the board. The child's task is to move the board to put the ball in the hole on the board. IoT functions to detect incoming balls and evaluate children's activities. The results of the activity will be stored on the IoT platform. The hardware components of this system are a microcontroller, IR obstacle avoidance sensor, buzzer, and drop box board. In this study, Wajdi et al. tested the system hardware's functionality and tested its performance based on two criteria: the accuracy of scoring children's activities and reading speed. In this study, Wajdi et al. proposed an IoT architecture based on four layers: constrained devices layer, console devices layer, communication network layer, and management services layer.

The previously mentioned studies published the integration of IoT with games to support motor development of preschoolers. Setiawan et al. build an IoT system integrated with Kobela aids at the elementary school level [27]. Kobela is an abbreviation of “Kotak Belajar Ajaib” (Magic Learning Box) which is used to support students in learning multiplication and division of numbers in mathematics. In this study, the authors tested the system based on hardware functionality and system performance: accuracy of children's activities and reading speed from sensors to the IoT platform. The integration of teaching aids in physics subjects was developed by Sakinah et al. and Fauzan et al. Sakinah et al. developed integrated props for Lorentz force practice with IoT

[28]. The system built by Sakinah et al. uses magnets and KY-024 module to detect magnetic fields. The data read by the KY-024 module is sent to the microcontroller to be sent to the system. The interaction of teachers and students to the system uses a website-based application. While in Fauzan et al.'s research, the author integrates IoT with free fall props to help students learn the concept of free fall [29]. These studies tested the system based on hardware functionality and system performance.

From the literature discussed, the development of IoT-based learning aids only focused on system development. In these studies, the researchers did not identify comprehensively: aspects of compliance with education 4.0, analysis of system functionality requirements, system performance analysis, system component analysis, system architecture design, system communication design, and system implementation and system evaluation. This study proposes a flexible and comprehensive model for developing an IoT-based teaching aid system that supports the implementation of education 4.0.

III. Proposed Method

The research method used in this study is the IoT development model proposed by Fahmideh and Zowghi. Fahmideh and Zowghi's method was modified in its sub-phases to suit the functionality and operation requirements of educational teaching aids. [30]. This model describes the process of developing an IoT application sequentially. Fig.1. shows the research method for the development of the IoT model. There are five phases in the development of IoT-based mathematical balance aids.

The first phase is the initialization phase. The purpose of this phase: to identify arithmetic operations learning problems that have solutions using IoT-based technology, conduct a literature study to identify development opportunities and current solutions, and examine the feasibility of environmental infrastructure to implement the system. The next phase is the analysis phase. This phase aims to identify functionality to support part of the education trend 4.0, namely free access, personalized learning, practice, and feedback. Designing the system architecture, then connecting between the hardware, designing the data flow diagrams, and designing the database are the goals of the design phase. Detailed explanations and components of the first, second, and third phases are explained in this second section.

After completing the design phase, the next phase is the implementation phase. In this phase, the implementation includes hardware implementation, software implementation, and system integration. The last phase in the research method is the evaluation phase. In the evaluation phase, there are two evaluations carried out. First, activities evaluate the achievement of system functionality that has been implemented. Second is the activity of assessing system performance based on aspects of system accuracy and delay. The two stages are explained in detail in Section III for the implementation stage and Section IV for the evaluation stage.

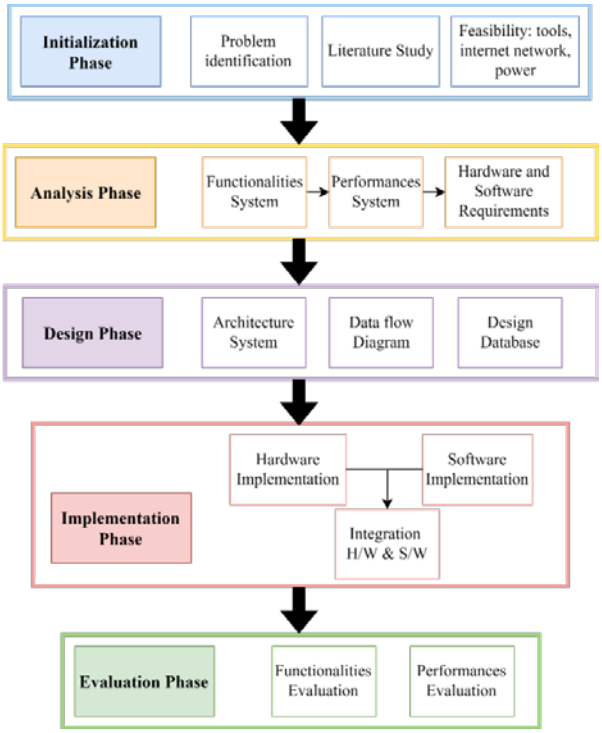


Fig. 1. Method for Integration Mathematical Balance aid with IoT based on Fahmideh and Zowghi [23].

A. Initialization Phase

The initialization phase is the first phase of this research. There are three activities in this phase. The first is to identify the problems of learning mathematics for elementary schools in Indonesia from the perspective of education 4.0. The next activity is to review the study literature to find out the position of this research in the scope of the application of IoT in the education sector. Section II carried out the literature review stage of this study. Checking the feasibility of supporting infrastructure to implement IoT-based solutions is the last activity in this phase. The authors identified problems from publications regarding the effect of using mathematical balance tools on learning outcomes and identification regarding the potential of IoT in providing opportunities for improving the quality of education and changing the character of education in the Industrial 4.0 era is conducted.

The last part of this phase is to check the feasibility of the availability of mathematical balance aids, hardware for IoT such as microcontrollers, and gyroscope sensors, the availability of internet networks in the classrooms, and the availability of electrical networks for the systems in the classrooms. After checking, the result is that all components and infrastructure are available to build an IoT-based mathematical balance teaching aid system.

B. Analysis Phase

The second phase in this research is the analysis phase which consists of three activities. The first activity is to define system functionalities based on research objectives, the second is to determine system performance, and the last activity is to determine hardware and software requirements following the specified functionality.

1) Functionalities system: System functionalities are determined based on research objectives. The objectives of this research are:

- to design and implement mathematical balance aids that students can use anywhere and anytime,
- to design and implement systems that can read and store activity results, and
- to design and implement systems that can provide real-time (i.e., the processing time in the system is less than 1 minute) and accurate feedback, and
- to be available from anywhere and anytime via the internet.

The functionality of this system is as follows:

- The system can be accessed using the internet network.
- The system has an application interface for users to interact with the system
- Mathematical balance aids integrated with IoT
- The system can send data on the results of student learning activities to the IoT platform
- The system has a database that stores student activities
- The system can identify users
- The system can evaluate activities accurately
- The system can display the evaluation results in real-time

2) Performances system: In this research, the evaluation of the system performance using accuracy parameters and delay parameters. The accuracy parameter ensures that student learning activities using mathematical balance aids are assessed correctly. This parameter is essential to determine the results of the feedback given to students. The delay parameter plays a role in ensuring it can meet Ambrose's feedback principle. Students need appropriate feedback immediately so that students can evaluate their activities.

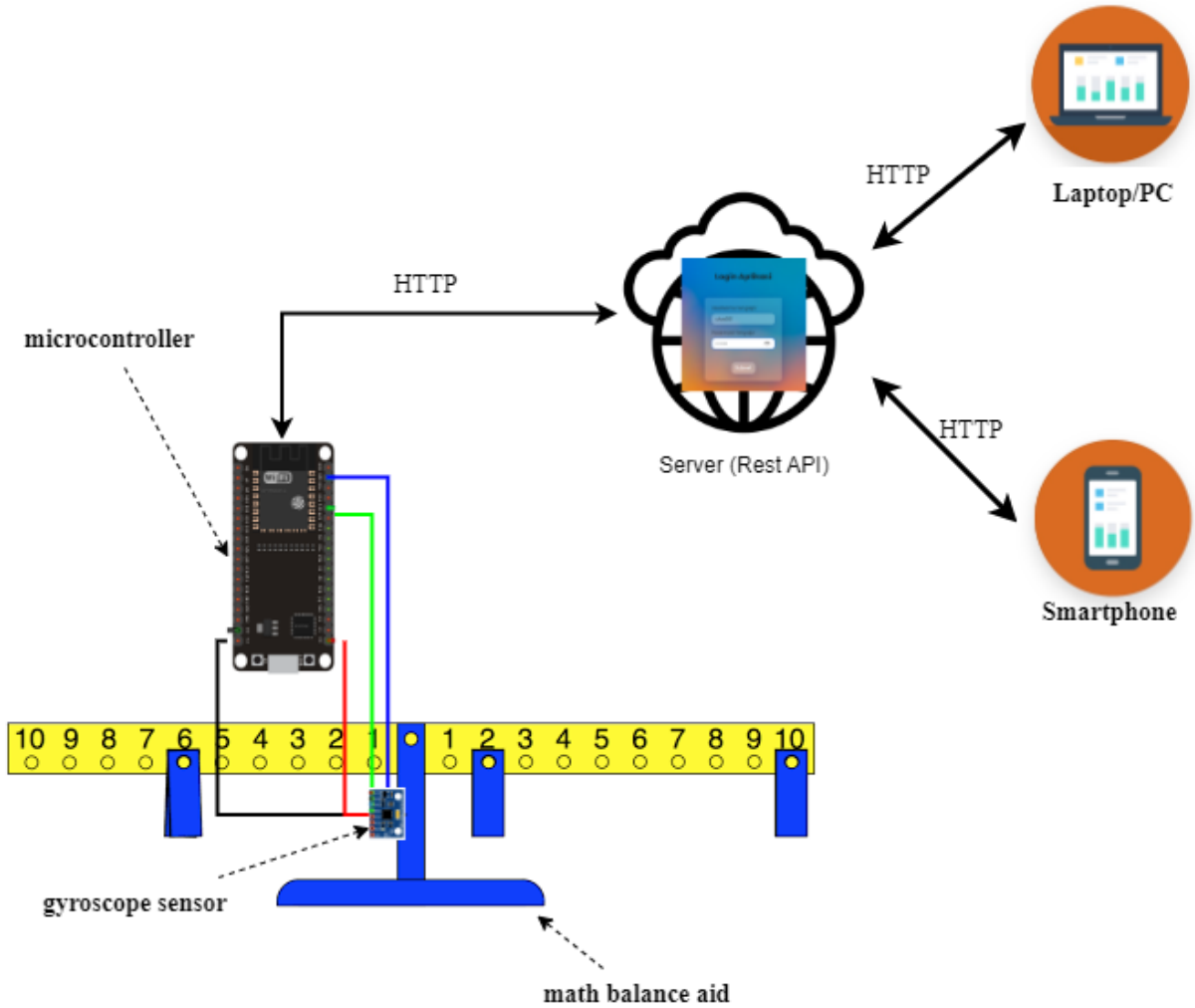


Fig. 2. Architecture System.

3) Hardware and software requirements: Based on the functionality and performance of the system that has been determined in the previous stage, at this stage, the requirements for hardware and software are defined. The hardware and software requirements to build an IoT-based mathematical balance teaching aid system are shown in TABLE I.

TABLE I. HARDWARE AND SOFTWARE REQUIREMENTS

Type	Component Name	Description
Hardware	Math balance Aids	Learning aids to support students learning about arithmetic operations
	Sensor gyroscope	Device for maintaining and measuring X-axis and Y-axis of math balance aids
	Microcontroller (with wire)	Board for receiving data and connecting math balance aids and gyroscope sensors to the application
	Battery or DC power supply	As a source of power to the system
	PC/Laptop/Smartphone	Device to access the application
Software	Web Application	An application that displays the login menu, calibrates the sensor, starts the class, and views the student's score.
	Database	An application for storing student data, grade data, class data, and teacher data
	Platform Communication	An application to connect the system to the internet network

C. Design Phase

The design phase consists of three component designs. The first component design is the system architecture. The system architecture shows how the components system and the communication between system components. The next component design is a data-flow diagram describing the developed application's data flow. Database design is the third component that shows the tables and the relationships between these tables.

1) Architecture system: Fig. 2 shows the system architecture for the IoT-based mathematical balance aids. In the system architecture, there are three components. The end node component consists of a mathematical balance teaching aid, gyroscope sensor, microcontroller, and wire. This component is used for student learning activities. The communication component consists of protocol communication (in this research use HTTP and TCP communication protocols). The website-based application is the last component, which functions to access and display student data, class data, grade data, and teacher/user data.

2) Data flow diagram: The data flow diagram shows the activity or operation of the system based on the data flow. The data flow diagram becomes a reference for designing the database. Fig. 3 shows the data flow diagram in this study, which consists of four functions: login, calibrate the sensor, start the class, and view student scores. In the login function to secure user/teacher data, the system provides additional functionality for credential data.

3) Database design: Database design refers to the data-flow diagram. This research designed the database to store data on student identity, student grades, classes, and users (teachers), and designing relationships between tables. The database label using the Indonesian language. In Indonesian, "murids" means student, "nilai" implies score, and "kelas" means class. There are two collections in the database design. The "murids" collection consists of two tables, namely the "murids" table and the "nilai" table.

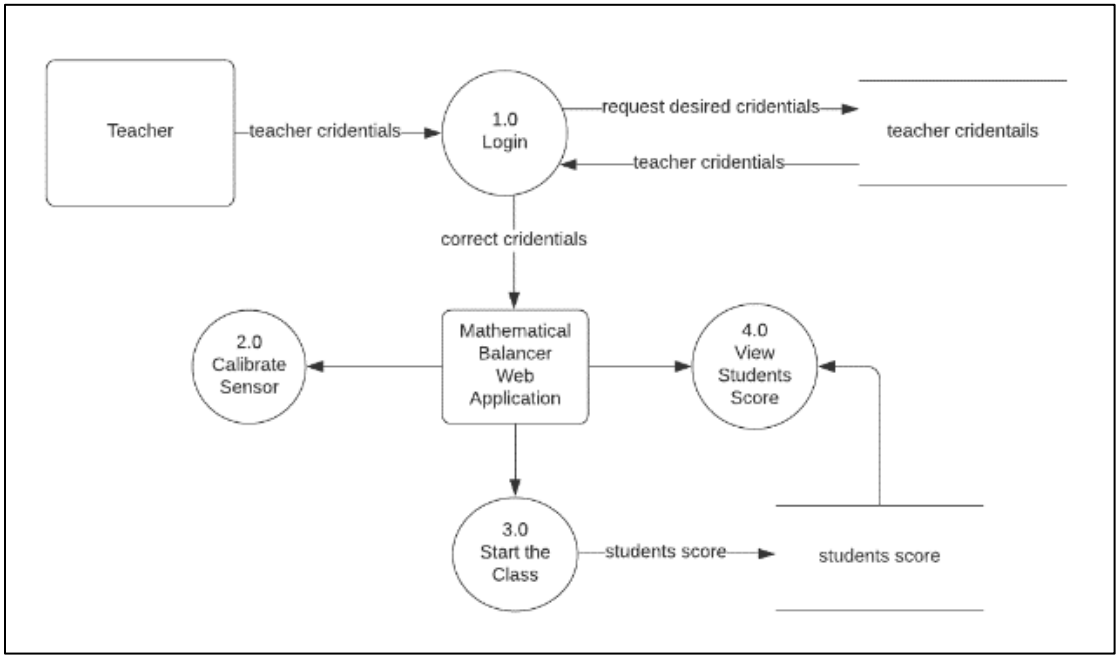


Fig. 3. Data-Flow Diagram.

The "murids" table stores student data such as name, age, and grades. The grades table has a date column, the number of grades (which contains the accumulated student activity grades) column, and the _id of the student column. In the "kelas" collection, there are two tables: the class table and the users' table. "Kelas" table store all registered student ids. The "kelas" table is related to users (or teachers). A user can relate to several different classes. Especially for table users, for data security, this research use salt and hash columns. The database design is shown in Fig. 4.

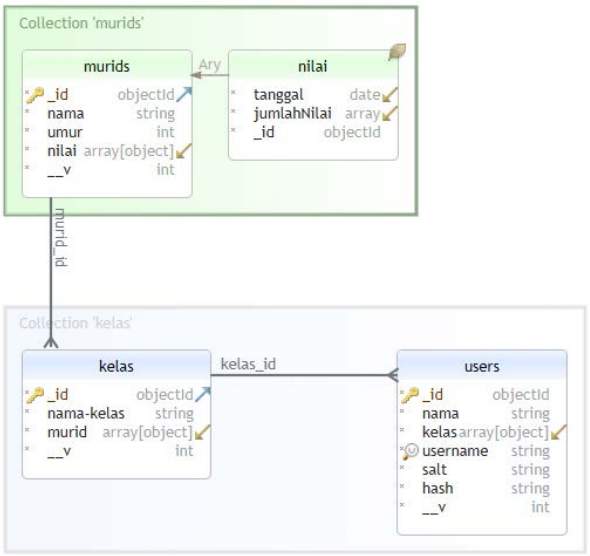


Fig. 4. Database Design.

D. Implementation Phase

The authors implement the system based on the design phase results in the implementation phase. There are three stages of implementation: hardware implementation, software implementation, and system integration. Hardware implementation describes how the author installs the hardware components of the IoT hardware and math balance aids.

The software implementation phase describes the stages of building software components. The stages of software implementation start from creating application interfaces, creating databases, and connecting application interfaces and databases. After the hardware and software components have been successfully implemented, the next step is integrating the two components. The discussion of system implementation is discussed in detail in Section III.

E. Evaluation Phase

The system's functionality and performance is tested in the evaluation phase. System functionality is tested by checking the suitability of system functionality with functionality in the analysis phase. Performances evaluation measures accuracy and delay. Accuracy is evaluated by comparing the student's answers with the expected answers, while for the delay, it will calculate the activity processing time. The author use test cases to evaluate system performance using 10, 100, and 1000 requests per second. A discussion of the evaluation phase is presented in Section IV.

IV. System Implementation

A. Hardware Implementation

Hardware implementation refers to the system architecture in Fig. 2. In implementing hardware components, the first step is to provide mathematical balance visual aids, gyroscope sensors, microcontrollers, wiring, power source support, and other required supporting components. The next step is to combine and configure the math balance aids with the gyroscope sensor. In a static position, all the gyroscope sensor's x, y, and z axes must have a value of 0. After the static position is obtained, the next step is to identify the pinout on the gyroscope sensor (in this study, the type of gyroscope sensor used is MPU6050), such as pins for power, pin for ground, pin for I2C communication, pin for I2C address, and an interrupt pin.

The next step is to prepare the Arduino IDE to program the microcontroller (in this study, the system uses ESP32). Prepare the libraries needed to integrate the ESP32 with the MPU6050 sensor module. After all the libraries have been prepared, the ESP32 and MPU6050 are integrated with the designed schematic diagram. The program is written on the ESP32 to read the axes and accelerometer from the gyroscope sensor. In this step, must ensure that under static conditions, the value of all axes is zero, and the acceleration value on the z-axis approaches the gravitational force value of 9.8 m/s2. For the x and y axes, the value is 0. If there is still a discrepancy in the value, it is necessary to calibrate the sensor.

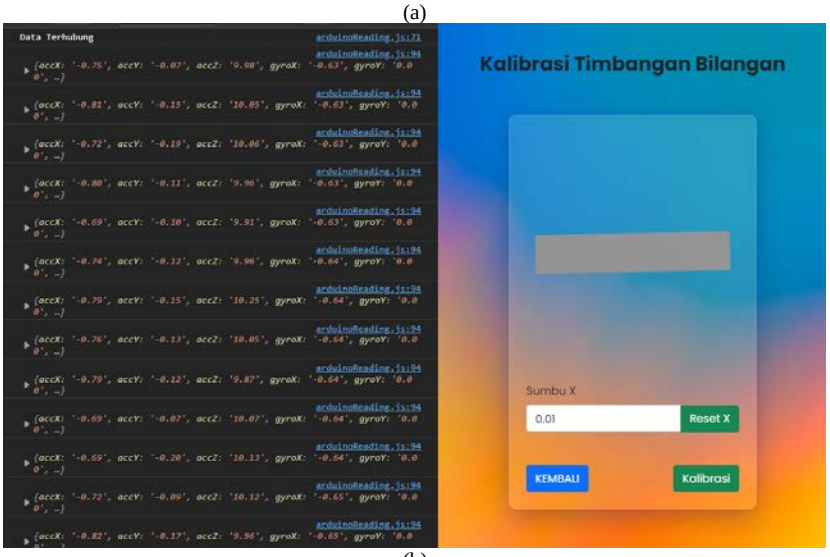
The mathematical balance aid is integrated with the gyroscope sensor. The system sends data to the internet network to be stored in the system database. Communication between the website and the microcontroller is two-way communication. The system allows setting the gyroscope sensor calibration via the website. The math balance aid is integrated with the gyroscope sensor to read the x, y, and z axes. The system assesses students' answers based on changes in the angle of the gyroscope sensor. The results of the system implementation are shown in Fig. 5.

Fig. 5. Implementation Hardware for Math Balance Aids Integrated with IoT.

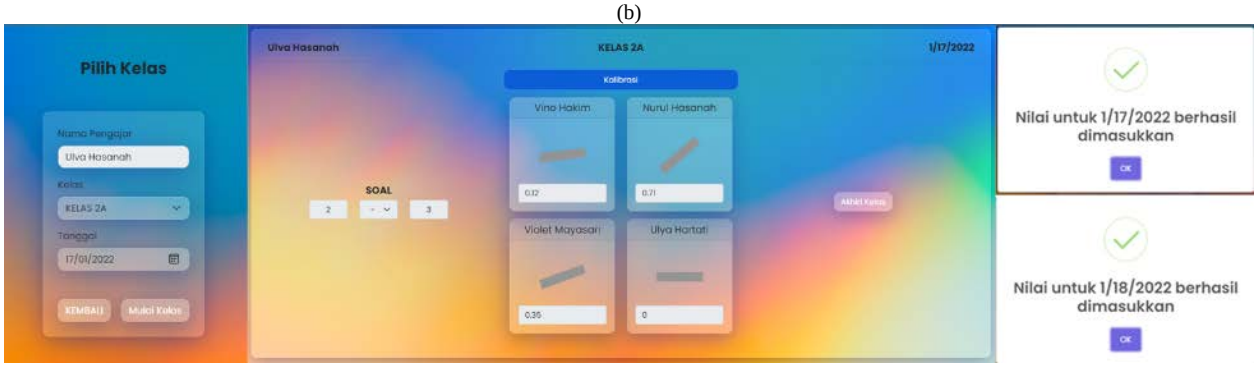
B. Software Implementation

The data flow diagram in Fig. 3 is the basis for implementing the software or application in this study. From the data flow diagram, the authors create an application interface design. Website-based applications allow users to access applications via the internet with various types of devices, such as personal computers, laptops, smartphones, or tablets. In this study, the application was built using node.js. The system implementation results are shown in Fig. 6 (a-d).

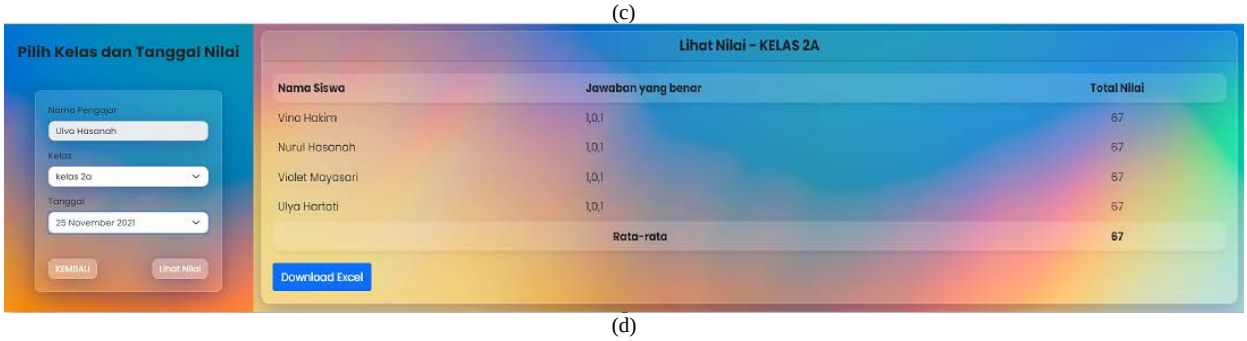
(a)



(b)



(c)



(d)

Fig. 6. (a) Login Page, (b) Calibration Page, (c) Start Class, (d) View Student Score.

Fig. 7. Database Implementation on MongoDB.

Fig. 6 (a) shows the display of the login page. Teachers use the login page to manage class activities for learning mathematical operations using IoT-based mathematical balance aids. Fig. 6 (b) shows a page for performing an application-based scale calibration. Before starting an activity, the teacher must calibrate the device to ensure the device is in a static state. After the class is started by the teacher, the display that appears in the application is Fig. 6 (c). At this stage, students carry out learning activities using IoT-based teaching aids. After completing the learning activity, Fig. 6 (d) shows the results obtained by the students. The database system is built based on the database design described at the design stage. The database system was built using the MongoDB application. Fig. 7 shows the implementation of the database according to the database design. After the interface and database have been successfully implemented, the next step is to integrate the two components.

C. Integration System

The microcontroller has a significant role in the integration process. The microcontroller stores the angle change data on the gyroscope sensor and sends it to a system that has been hosted online. So that student activity data can be accessed by users using various devices via the internet. The implementation phase is complete when all data can be read, stored, processed, and accessed according to the system design.

V. Evaluation and Discussion

This section discusses the evaluation of the system applied in the previous section. This section discusses the potential for improving the systems that have been built and the opportunities for developing IoT-based learning systems in education in the future.

A. Evaluation Phase

1) Functionaliy evaluation: This stage evaluates the functionality that has been defined in Section 2.2.1. The evaluation was conducted by measuring the achievement of functionality in the system. The summary of achieving the functionality defined by the system functionality is shown in Table II.

Testing of website functionality and website menus is carried out using the test case method. The test results for each website's functionality and website menu are shown in Table III.

TABLE II. SYSTEM FUNCTIONALITIES

No	Functionalities	Implementation
F1	The system can be accessed using the internet network.	The authors build a website-based software system so that it can be accessed via the internet and can be accessed with various devices such as personal computers, laptops, or smartphones. The interface of the system is shown in Fig. 6.
F2	The system has an application interface for users to interact with the system	The system has an interface that allows users to perform interactions such as logging in, calibrating teaching aids, starting class activities, and viewing student grades. As shown in TABLE III.
F3	Mathematical balance aids integrated with IoT	Mathematical balance aids are successfully integrated with IoT. System integration with IoT can be seen from the system architecture in Fig. 2. and the system implementation in Fig. 5.
F4	The system can send data on the results of student learning activities to the IoT platform	The results of arithmetic operations learning activities using IoT-based balance aids can be sent to the IoT platform as shown in Fig. 6(d).
F5	The system has a database that stores student activities	The authors implement the database based on the database design in Fig. 4., and the results of the database implementation are shown in Fig. 7.
F6	The system is able to identify users	The system checks the login process to ensure that only authorized users can log in to the system and send error message notification. An example of a login process that does not meet system authentication is shown in the figure
F7	The system can evaluate activities accurately	The system can process the input into the output of the assessment system, as shown in TABLE IV.
F8	The system is able to display the evaluation results in real-time	The system can process the assessment results in real-time with the specified number of requests, as shown in TABLE V.

TABLE III. TEST CASE FOR WEBSITE PAGE FUNCTIONALITIES AND THE TEST RESULTS

No	Functionality	Test Case	State
1	Login	Verify credentials	succeed
2		Sending error message if the credentials are incorrect	succeed
3	Calibrate Sensor	Checking connection from the microcontroller	succeed
4		Sending error message if microcontroller is not connected properly	succeed
5		Generate the angle of inclination value from the microcontroller	succeed
6		Calibrate the sensor and make sure the angle of inclination is 0 radian	succeed
7	Start the Class	The teacher can choose a class to teach	succeed
8		The class name provided by the system is the same as the teacher signed at	succeed
9		The teacher able to start the class after providing class name and the current date	succeed
10		The system only able to process number from zero to ten	succeed
11		The angle of inclination from each student microcontroller are displayed correctly	succeed
12		The system sends the angle of inclination value to database	succeed
13		The system would show error message if the data failed to be sent to database	succeed
14	View Students Score	The teacher can choose a class where he teach	succeed
15		The class name provided by the system is the same as the teacher signed at	succeed
16		The system can process and evaluate the final students score correctly	succeed
17		The system can export the scores to a .xls file	succeed

2) Performance evaluation: There are two aspects to evaluating system performance. The first is to test and evaluate the accuracy between the results obtained from the angle of inclination of the mathematical balance aids with the expected result. The tolerance limit for the angle of inclination in this study was 0.14 degrees. For answers with angles below or equal to 0.14 degrees, the value is 1. In contrast, at angles that exceed the limit, the value is 0. The results of the device testing for five types of questions given to eight children got the results as shown in TABLE IV. For question number one, there is a question of adding the number one to number 2. On one side of the balance tool arm will be given a load on the number 1 and number 2. The student's task is to put the burden on a certain number that is the answer to the question. When a student puts a load on a specific number, and the inclination angle exceeds 0.14 degrees, the student's response is worth 0. The result of this answer is then compared with the expected result. If both the result and expected result are the same, then the tool assesses the answer accurately. For example, a student named Vino Hakim gave an answer that caused an angle of inclination of 0.04 degrees so that the result is worth 1, when compared to the expected result, which is worth 1, so the authors give a valid status.

TABLE IV. ACCURACY OF IOT-BASED MATH BALANCE AIDS

Question	Student	Angle of Inclination (degrees)	Result	Expected Result	Status
1 + 2	Vino Hakim	0.04	1	1	Valid
	Nurul Hasanah	0.09	1	1	Valid
	Violet Mayasari	0.2	0	0	Valid
	Ulya Hartati	0.13	1	1	Valid
4 - 3	Endah Riyanti	0.02	1	1	Valid
	Adinata Mansur	0	1	1	Valid
	Kuncara Januar	0.07	1	1	Valid
	Ilsa Hartati	0.14	1	1	Valid
3 x 2	Vino Hakim	0.17	0	0	Valid
	Nurul Hasanah	0.15	0	0	Valid
	Violet Mayasari	0.31	0	0	Valid
	Ulya Hartati	0.07	1	1	Valid
4 ÷ 2	Endah Riyanti	0.02	1	1	Valid
	Adinata Mansur	0.32	0	0	Valid
	Kuncara Januar	0.12	1	1	Valid
	Ilsa Hartati	0.02	1	1	Valid

The following evaluation is to measure system delay. In this study, the system delay is calculated from the time of reading data from the gyroscope sensor to the microcontroller to accessing the user's device using the internet network. This study uses a simple architecture from one mathematical balance aid to multiple user devices to measure the delay time. In the microcontroller, the authors set the data reading from the gyroscope sensor to 10 milliseconds, and the connection delay from the microcontroller to the WIFI is one second. The speed of WIFI is approximately 10 Mbps. This configuration sends one hundred data from the sensor to the website in less than two seconds. The next is the measurement of a delay from the website to the user's device. By using a web performance test, the test results for 10, 100, and 1000 requests to the system resulted in 100% completed requests and with a total time of fewer than 4 seconds, as shown in TABLE V. Total time required to read a hundred data from the sensors to the application with 1000 requests in less than 10 seconds.

TABLE V. PERFORMANCE TEST RESULTS

Requests	Total Errors	Completed Request	Request/s (per second)	Total Time
10	0	10	97	0.10s
100	0	100	216	0.46s
1000	0	1000	276	3.6s

B. Discussion

The modification of the IoT development method proposed by Fahmideh and Zowghi in this study provides a comprehensive guide for developing IoT-based teaching aids. This study used the development method to build math balance aids to help students learn arithmetic operations. The development of an IoT system based on mathematical balance aids with the proposed method shows the successful integration of mathematical balance aids and IoT functionality to achieve the three characteristics of education 4.0: remote access, personalized learning, and practice and feedback.

Fahmideh and Zowghi's IoT development method does not discuss the specifics of IoT implementation in certain fields. The modification of Fahmideh and Zowghi's method in this study is specifically intended for the field of education. The method proposed in this study provides guidance in identifying the need for IoT integration with education 4.0. In the future, the method in this research can be used to implement IoT with teaching aids at various education levels and various subjects or courses.

Although this study succeeded in building a prototype of an IoT-based math balance aid in accordance with each phase of the proposed method, several limitations have not been resolved in this study. From the aspect of durability and hardware packaging, the system still has to be developed to be more robust. The gyroscope sensor must be corrected in several tests to obtain valid results. In future, research that examines the durability and packaging of the system will be necessary. From the software aspect, applications must be developed by considering aspects of the user interface (UI) and user experience (UX). There is potential for UI/UX development to develop applications for IoT systems by considering user personas and human factor considerations so that the resulting application is an application that is easy, interactive, and convenient for users (teachers, students, and education staff) [31].

Furthermore, from the educational aspect, application development is integrated with existing educational applications to provide comprehensive feedback data. Integrating the system with data science or artificial intelligence to improve the quality of feedback based on student learning activity data is a future research opportunity [32]. Gamification can be combined with the system to increase student motivation or engagement. There are still many opportunities to conduct research on gamification in IoT, especially for education and industry 4.0 [33]. This research is still being tested on a laboratory scale. It needs to be tested in the classroom to determine technology acceptance by teachers, students, or academic staff. This research also provides an opportunity to examine the impact of the system on student learning outcomes. Another aspect that is not discussed in this study is the aspect of data security and privacy. Based on various studies, the security and privacy issue in IoT is one of the biggest challenges in IoT implementation [11], [34], [35]. Activity result data is published on the internet network, which has an impact on the emergence of opportunities for theft or use of data by unauthorized parties.

VI. CONCLUSION

In this study, an IoT-based mathematical balance tool has been successfully developed to support arithmetic operational learning for elementary school students in Indonesia. The purpose of implementing IoT in teaching aids in this study is to design education that supports education 4.0. This study proposes an IoT development method modified from the method proposed by Fahmideh and Zowghi. The IoT development method consists of five stages. The proposed IoT development method defines each stage comprehensively. The first stage is the initialization stage, which identifies research problems, conducts related literature studies, and identifies the feasibility of supporting infrastructure. The second stage is the analysis stage with the following objectives: defining system functionality according to research objectives and defining system performance based on system accuracy and delay. The third stage is the design stage, to determine the requirements for hardware and software. The system architecture, data flow diagrams, and database design are also designed at this stage. The results of the third stage become a reference for the next stage, namely the implementation stage. The evaluation of the system implementation showed that all functionality was successfully implemented with 100% accuracy for reading, recording, and evaluating student activities. While the results of the evaluation of the system performance show system delay for 1000 requests is less than 10 seconds.

ACKNOWLEDGMENT

This work is supported by internal funding from Telkom University.

REFERENCES

[1] C. D. Jerald, “Defining a 21st century education,” Center for Public education, 2009.

[2] R. Drath and A. Horch, “Industrie 4.0: Hit or Hype? [Industry Forum],” IEEE Ind. Electron. Mag., vol. 8, no. 2, pp. 56–58, Jun. 2014.

[3] P. Fisk, “Education 4.0 … the future of learning will be dramatically different, in school and throughout life,” www.peterfisk.com, 2017. [Online]. Available: https://www.peterfisk.com/2017/01/future-education-young-everyone-taught-together/. [Accessed: 11-Mar-2022].

[4] J. Miranda et al., “The core components of education 4.0 in higher education: Three case studies in engineering education,” Comput. Electr. Eng., vol. 93, p. 107278, Jul. 2021.

[5] Oliver, “The role of ICT in higher education for the 21st century: ICT as a change agent for education,” Retrieved April, 2002.

[6] Z. I. Ciroma, “ICT and Education: Issues and Challenges,” Mediterr. J. Soc. Sci., vol. 5, no. 26, pp. 98–98, Dec. 2014.

[7] S. Talebian, H. M. Mohammadi, and A. Rezvanfar, “Information and Communication Technology (ICT) in Higher Education: Advantages, Disadvantages, Conveniences and Limitations of Applying E-learning to Agricultural Students in Iran,” Procedia - Social and Behavioral Sciences, vol. 152, pp. 300–305, Oct. 2014.

[8] B. Dorsemaine, J.-P. Gaulier, J.-P. Wary, N. Kheir, and P. Urien, “Internet of Things: A Definition & Taxonomy,” in 2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies, 2015, pp. 72–77.

[9] R. Minerva, A. Biru, and D. Rotondi, “Towards a definition of the Internet of Things (IoT),” IEEE Internet Initiative, vol. 1, no. 1, pp. 1–86, 2015.

[10] González García and Núñez Valdéz, “A review of artificial intelligence in the internet of things,” Artif. Intell. Appl., 2019.

[11] M. Kassab, J. DeFranco, and P. Laplante, “A systematic literature review on Internet of things in education: Benefits and challenges,” J. Comput. Assist. Learn., vol. 36, no. 2, pp. 115–127, Apr. 2020.

[12] S. A. Ambrose, M. W. Bridges, M. DiPietro, M. C. Lovett, and M. K. Norman, How Learning Works: Seven Research-Based Principles for Smart Teaching. John Wiley & Sons, 2010.

[13] Saeed, Munir, and Shah, “Usage Of Internet Of Things (IoT) Technology In The Higher Education Sector,” J. At. Mol. Phys., 2021.

[14] Cornelius, Michael, and Ed, Teaching Mathematics. Nichols Publishing Company, P.O. Box 96, New York, NY 10024 ($15. paper copy, $27.50 cloth copy); Croom Helm Ltd., 2-10 St. John’s Road, London SW11, England., 1982.

[15] Little, “Teaching mathematics: Issues and solutions,” Teach. Except. Child. Plus, 2009.

[16] Suherman, “Strategi pembelajaran matematika kontemporer,” Bandung: Jica, 2003.

[17] Miyarso, “Pengembangan Alat Peraga Timbangan Untuk Mengoptimalkan Belajar Hitung Bagi Siswa SD,” Majalah Ilmiah Pembelajaran, 2011.

[18] A. Sukesih, “Peningkatan Hasil Belajar Siswa Pada Mata Pelajaran Matematika Melalui Media Neraca Bilangan Di Kelas II Sekolah Dasar Islam Terpadu Aziziyyah Kecamatan Tampan Pekanbaru,” skripsi, Universitas Islam Negeri Sultan Syarif Kasim Riau, 2017.

[19] R. N. Sari, “Pengaruh Penggunaan Alat Peraga Timbangan Bilangan Terhadap Pemahaman Konsep Perkalian Di Kelas II SDI Al Azhar 15 Pamulang,” Jakarta: FITK UIN Syarif Hidayatullah Jakarta, 2018.

[20] Tryaji and Listyarini, “Penerapan Pendekatan Discovery Learning Berbantu Neraca Bilangan Terhadap Hasil Belajar Matematika Kelas III Materi,” Proceeding TEAM, 2018.

[21] T. K. Rachmawati, E. Farlina, W. Setya, and W. A. Tutut, “Penggunaan Alat Peraga Timbangan pada Materi Bilangan Asli dan Kesetimbangan,” J-ABDIPAMAS (Jurnal Pengabdian Kepada Masyarakat), vol. 3, no. 2, pp. 63–72, Oct. 2019.

[22] A. W. Putri and D. Damri, “Efektivitas Penggunaan Media Neraca Bilangan Untuk Meningkatkan Pemahaman Konsep Operasi Perkalian Bagi Siswa Tunagrahita Ringan,” Jurnal Cendekia : Jurnal Pendidikan Matematika, vol. 4, no. 2, pp. 1164–1170, Nov. 2020.

[23] S. Shonia and N. A. Suwastika, “Bag Toss Game based on Internet of Education Things (IoET) for the Development of Fine Motor Stimulation in Children 5-6 Years Old,” EMITTER International Journal of Engineering Technology, 2020.

[24] [24] I. N. Rahmanto, N. A. Suwastika, and R. Yasirandi, “How Can IoT Applicable to Practice Gross Motor Skill Through Hopscotch Game?,” Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 4, no. 3, pp. 584–590, Jun. 2020.

[25] R. K. Jati, N. A. Suwastika, and R. Yasirandi, “Hopscotch game to support stimulus in children’s gross motor skill using IoT,” Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control, vol. 5, no. 4, pp. 277–290, Nov. 2020.

[26] H. Wajdi and N. A. Suwastika, “IoT architecture that supports the stimulation of gross motor development in children aged 5-6 years using drop box game,” Register: Jurnal Ilmiah, 2020.

[27] M. I. Setiawan, N. A. Suwastika, and S. Prabowo, “IoT-Based Kobela Teaching Aid for Mathematics Learning Multiplication and Division Materials for Grade II Elementary School Students,” JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 5, no. 3, pp. 1142–1149, Jul. 2021.

[28] H. R. Sakinah, N. A. Suwastika, M. Al Makky, and Q. Qonita, “Lorentz Force Experiment Prop based on IoT (E-Lorentz) to Support the Learning Process of Physics Subject for 9th Grade,” JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 5, no. 4, pp. 1283–1291, Oct. 2021.

[29] M. N. Fauzan, N. A. Suwastika, and E. M. Jadied, “Internet of Things (IoT) Based Free Fall Motion Instructions in Physics Subjects for Class X Students,” JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 6, no. 2, pp. 876–886, Apr. 2022.

[30] M. Fahmideh and D. Zowghi, “An exploration of IoT platform development,” Inf. Syst., vol. 87, p. 101409, Jan. 2020.

[31] P. Abichandani, V. Sivakumar, D. Lobo, C. Iaboni, and P. Shekhar, “Internet-of-Things Curriculum, Pedagogy, and Assessment for STEM Education: A Review of Literature,” IEEE Access, vol. 10, pp. 38351–38369, 2022.

[32] J. J. E. McBroom, “Data Science to Improve Feedback, Understand Student Behaviour and Address Equity Issues in Computer Programming Education Using Data from Large-Scale Courses,” ses.library.usyd.edu.au, 2021.

[33] R. Xiao, Z. Wu, and J. Hamari, “Internet-of-Gamification: A Review of Literature on IoT-enabled Gamification for User Engagement,” International Journal of Human–Computer Interaction, pp. 1–25, Dec. 2021.

[34] D. D. Ramlowat and B. K. Pattanayak, “Exploring the Internet of Things (IoT) in Education: A Review,” in Information Systems Design and Intelligent Applications, 2019, pp. 245–255.

[35] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, “A Survey on Security and Privacy Issues in Internet-of-Things,” IEEE Internet of Things Journal, vol. 4, no. 5, pp. 1250–1258, Oct. 2017.

An Efficient Patient Activity Recognition using LSTM Network and High-Fidelity Body Pose Tracking

Thanh-Nghi Doan
Faculty of Information Technology, An Giang University
Vietnam National University Ho Chi Minh City
An Giang, Vietnam

***Abstract*—The need for healthcare services is growing, particularly in light of the COVID-19 epidemic's convoluted trajectory. This causes overcrowding in medical facilities, making it difficult to manage, treat, and monitor patients' health. Therefore, a method to remotely observe the patient's behavior is required, to aid in early warning and treatment, and to reduce the need for hospitalization for patients with minor diseases. This paper proposes a new real-time smart camera system to monitor, recognize and warn the patient's abnormal actions remotely with reasonable cost and easy to deploy in practice. The key benefit of the proposed methods is that patient actions may be detected without the usage of ambient sensors by employing pictures from a regular video camera. It carries out the detection using high-fidelity human body pose tracking with MediaPipe Pose. Then, the Raspberry Pi 4 device and the LSTM network are used for remote monitoring and real-time classification of patient actions. The test dataset is built from reality and reuses the existing datasets. Our system has been evaluated and tested in practice with over 96.84% accuracy, runs at over 30 frames per second, suitable for real-time execution on mobile devices with limited hardware configuration.**

***Keywords*—Human body pose tracking; LSTM; raspberry Pi 4; patient monitoring system**

I. Introduction

The development of efficient and reliable remote patient action recognition systems has been receiving much attention from the scientific research community. The benefits of patient monitoring from a distance include the ability to detect illnesses early and in real time, monitor patients continuously, stop illnesses from getting worse and prevent untimely deaths, lower hospitalization costs, fewer hospitalizations, and more accurate readings while still allowing patients to go about their daily lives normally. By using communication technology, emergency medical services, care for patients with mobility issues, emergency care for injuries sustained in traffic accidents and other types of accidents, and non-invasive medical interventions, healthcare services are made more efficient. In recent years, action recognition methods have focused heavily on the use of image and video analysis technologies. There are different definitions of action recognition presented in the study by Herath et al. [1]. The rapid development of smart devices and deep learning techniques have spurred the development of action recognition systems. These techniques have been widely applied in life such as entertainment, monitoring and human health care [2]. However, according to the survey by Szegedy et al. [3], the identification of complex and specific actions is still a big challenge to study. The articles [4], [5] presented a comprehensive review of fall detection systems and remote patient action recognition. Researchers have built a large variety of systems that can operate with the many technologies used to monitor patient behavior. These systems are broadly characterized as wearable, ambient, and computer-vision-based [6]. Researchers have created a vast array of systems that can work with the various technologies that are used to track patient activity. In general, these systems can be divided into wearable, ambient, and computer-vision-based ones [6].

The first block, wearable systems, includes sensors carried by the monitored individual. This set of systems employs a wide range of technologies, including accelerometers, pressure sensors, inclinometers, gyroscopes, and microphones, among others. The authors of [7] carefully evaluate and study these systems. The study attempts to assess the state of the art in such monitoring, both in terms of the most commonly used sensor technologies and their placement on the human body. These techniques, however, have the problem of requiring the devices to be put on the individuals' bodies. Because this sort of sensor must be worn continually, it might be unpleasant and not always viable [8]. The second block contains devices with pressure, acoustic, infrared, and radio-frequency sensors that are positioned around the monitored individual [9], [10]. However, the expense of installing these systems is very costly, and they are only appropriate for specialist patient care rooms, making them difficult to employ in everyday living at home. The last block, which is the focus of this research, groups systems capable of identifying human recognition using image-based computer vision. In recent years, convolutional neural networks (CNN) [11] have been widely used in image classification problems in many fields. Due to CNN's superior performance [12], many studies have started to use CNN for video classification. Long Short-Term Memory (LSTM) neural networks and conventional CNNs, or a combination of the two, have both been shown to perform well in human action recognition (HAR). In which CNN has been used to analyze sensor data for HAR with exceptional results [13]. Previous studies have proposed to supplement the feature vector extracted by CNN with some statistical features [14]. Aviléz-Cruz et al. [15] have developed a three-input CNN model to

recognize six human actions. The usefulness of LSTM networks for HAR has been demonstrated by additional studies as well [16]. Finally, several studies have suggested enhancing CNN with LSTM layers [17]. Recently, the article [18] proposed a network model that combines LSTM, MobileNetV2 and Raspberry Pi 4 in remote patient action monitoring and identification. However, these methods in the last block have the drawback of making it difficult to distinguish between closely related actions, such as waving and clapping, walking and running. Furthermore, training network models takes a long time due to direct learning of data from video frames, where CNN models are used to extract features from video frames. Because of this, these methods require extensive hardware configurations and have slow response times.

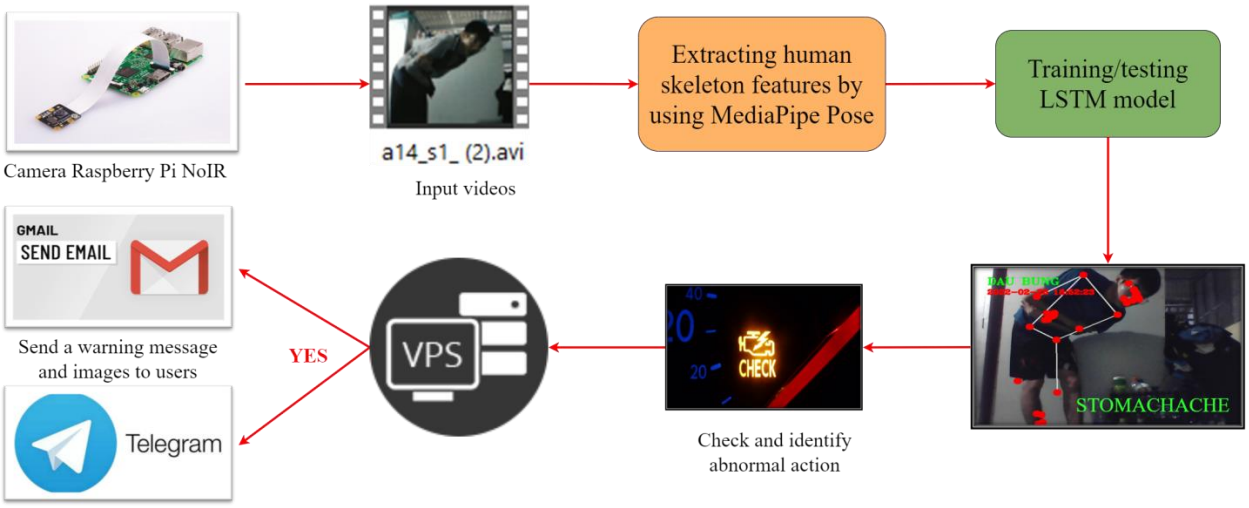


Fig. 1. Overview of our Proposed System for Monitoring Patient Healthy.

On the other hand, studies on the skeleton features extraction method for patient action identification [19], [20], and [21] have demonstrated a number of benefits that can get around the issues raised above. As a result, this paper proposes a new system for real-time identification of patient actions that is low in cost, efficient, has a fast response time, and is simple to install and implement in practice using hardware devices with limited configuration. The main contributions of the paper include:

- A novel method for patient action recognition using MediaPipe Pose framework [22], LSTM network model and Raspberry Pi 4 device [23].
- A new dataset was created using 541 videos of 16 different types of patient actions that were preprocessed and labeled in accordance with benchmark dataset standards.
- An action recognition system that is fully developed, user-friendly, and capable of continuous monitoring and alerting of abnormal human activity of the patient at home.

The rest of the article is arranged as follows. Section II describes the materials and methods used to describe overview of our system, data collection, patient action recognition model, Raspberry Pi and Camera Raspberry Pi NoIR . The experimental results and discussion are reported in Section III. Section IV presents the conclusions, limitations, and recommendations for future research.

II. Materials and Methods

A. Overview of our System

The overview of the proposed system for the remote patient monitoring camera system is shown in Fig. 1. In which the Raspberry Pi 4 Camera Module NoIR [23] is used to continuously monitor the patient's activities at home in real time. The generated video sequence is recognized and labeled in real time using the MediaPipe Pose framework [24] and the LSTM network model [25] trained and saved on a Raspberry Pi 4 device. If the patient's actions are considered to be abnormal, there is a health problem, and the system will immediately send a warning message, along with a photo of the odd activity, to the patient's relatives via email and the *Telegram* messaging application. The labeled videos are then saved on a virtual server on a regular schedule. Videos labeled as abnormal actions will be kept on the server for a long time, whereas normal actions will be kept for a short time and deleted after a certain period of time to save storage space. The algorithm to recognize real-time patient actions in videos and send warning messages with images of abnormal actions to users is summarized and shown in Fig. 2.

B. Data Collection

There are many published datasets on human action recognition such as ActivityNet [26], Kinetics [27], UCF101 [28], HMDB51 [29], STAIR-Actions [30], KARD [31], and NTU RGB+D [32]. However, these datasets do not include recordings of patient activities, and there is presently no published official benchmark dataset for these types of activities. Therefore, this study self-constructed a new dataset on patient actions to test our proposed approach. This dataset combines existing data with data generated by us from the actual world. A summary of this dataset is presented in Table I.

```
BEGIN
    Input: LSTM model, patient body skeleton points
    Label = "Nomarl action"
    The model predict the result based on the skeleton points
    IF Result = 1:
        Label = 'Hand swing'
    ELIF Result =2:
        Label = 'Hand clap'
     .
     .
     .
    ELIF Result =14:
    {
        Label = 'Stomachache'; Save the stomachache images
        Send emails, messages, images via Gmail and Telegram
    }
     .
     .
     .
    ELSE
        Label = 'Normal action'
    RETURN: Label
END
```

Fig. 2. The Algorithm Processes Skeleton Point Data, Returns Results and Sends Notification Messages to GMAIL, TELEGRAM.

TABLE I. A SUMMARY OF OUR DATASET ITH 16 TYPES OF PATIENT ACTIONS

ID	Description	Avg frame	Frame		Frame/s	Number of videos
			Width	Height		
a01	Hand swing	2500	640	480	25	81
a02	Hand clap	2500	640	480	25	23
a03	Body swing	2500	640	480	25	24
a04	Drink	2500	640	480	25	51
a05	Sit down	2500	640	480	25	34
a06	Stand up	2500	640	480	25	33
a07	Walking	2500	640	480	25	55
a08	Side kick	2500	640	480	25	66
a09	Phone call	2500	640	480	25	34
a10	Hand pain	2500	640	480	25	15
a11	Leg pain	2500	640	480	25	22
a12	Headache	2500	640	480	25	25
a13	Neck pain	2500	640	480	25	25
a14	Stomachache	2500	640	480	25	22
a15	Backache	2500	640	480	25	19
a16	Fall down	2500	640	480	25	12
Total number of videos						**541**

Due to time and staffing restrictions, we could only create a test dataset with 16 examples of the patient's actions. These actions are collected and separated into two groups: (i) the patient's normal actions (shown in the blue bounding box of Fig. 3) and (ii) the patient's abnormal actions (shown in the red bounding box of Fig. 3. This dataset includes four actions taken from the KARD dataset [31] and 12 actions we independently created by recording patient action video clips in the real experimental setting. KARD is a dataset that includes 18 different types of indoor daily activities with a resolution of 640x480 and reasonably clear action gestures. Consequently, they can be utilized to develop and evaluate a patient health monitoring system at home. However, only four action classes that are appropriate for this problem are used in this study: sit down, stand up, side kick, and phone call. We recruited volunteers to carry out 12 distinct types of actions for fact-generated data. Each type of action was performed three times and video recorded for three seconds each, using a Webcam HD 720p with the detailed specifications shown in Table II.

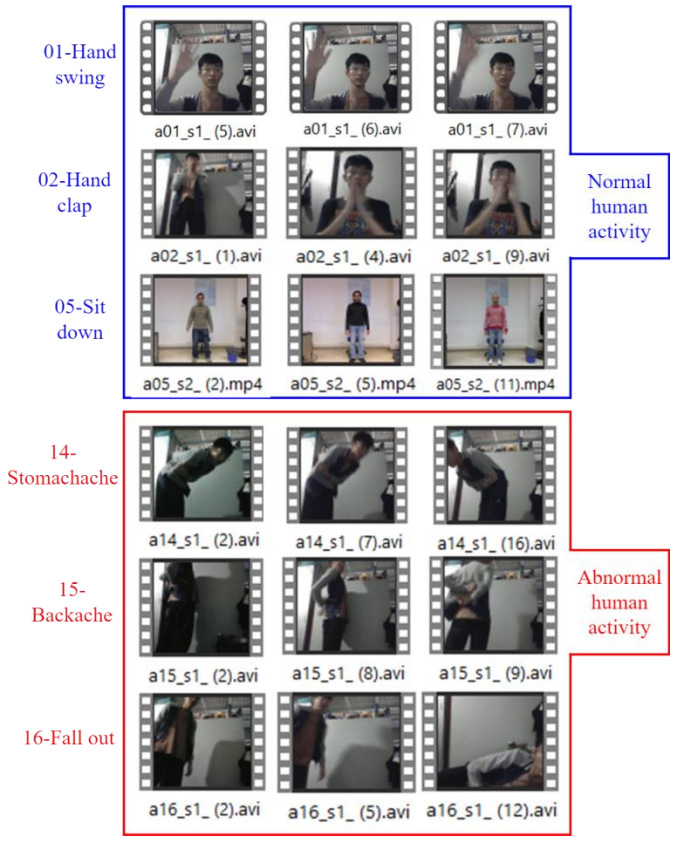


Fig. 3. Video Samples of our Patient Activities Datasets.

TABLE II. THE DETAILED SPECIFICATIONS OF WEBCAM HD 720P.

Specifications	Value
Camera	HD webcam 720p
Resolution	800x600
FPS	25 frames/second
Camera color	Color
Flash mode	No
Focus type	Fixed focus
Video format	AVI

The total number of videos we have collected is over 700 videos. These videos are then preprocessed, and the videos that do not meet the quality requirements are removed, yielding a video dataset of 541 files. The total size of the video dataset is 241 MB in which, the number of videos of each action type ranges from 12 to 81 videos, as shown in Table I. Each video is

shot at a frame rate of 25 FPS. This dataset is annotated in order of videos in each folder according to each action type, i.e. a01_s1_(1).mp4, a01_s01_(2).mp4,…, a02_s1_(1).mp4, a02_s01_(2).mp4,… equivalent to actions labeled as a01 (Hand swing), a02 (Hand clap), a03 (Body swing), a04 (Drink), a05 (Sit down), a06 (Stand up), a07 (Walking)), a08 (Side kick), a09 (Phone call), a10 (Hand pain), a11 (Leg pain), a12 (Headache), a13 (Neck pain), a14 (Stomachache), a15 (Backache) and a16 (Fall down). Some sample videos of the dataset consisting of 09 normal actions and 07 abnormal actions of the patient, are presented as shown in Fig. 3.

C. Patient Action Recognition Model

1) Extract Human Body Features with MediaPipe Pose:

MediaPipe Pose is a machine learning solution for high-fidelity body pose monitoring that uses the BlazePose research [22] to infer 33 3D landmarks and a background segmentation mask on the entire body from RGB video frames. The network can generate 33 body keypoints for a single human during inference and performs at over 30 frames per second on a Pixel 2 phone. Therefore, it is well suited to real-time applications such as fitness tracking and sign language recognition. The benefit of this skeletal feature extraction method is its real-time speed, fast response time, and good results even with low-quality and low-resolution video clips, independent of ambient variables such as light, shadow, and the ability to identify many objects at the same time.

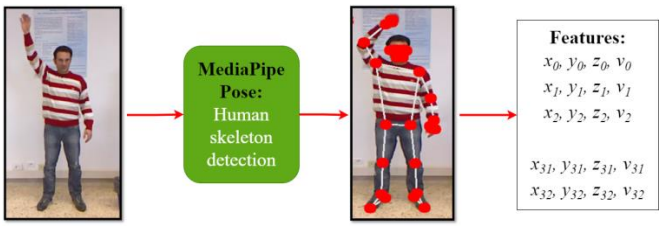


Fig. 4. Extracting Human Skeleton Features using MediaPipe Pose [24].

In this study, skeleton features are extracted from each frame of video using the algorithm described in the paper [22]. As shown in Fig. 4, the result of extracting each video frame is 33 skeleton points corresponding to 33 coordinates (*x*, *y*, *z*) and a visibility value *v* numbered from 0 to 32. Each skeleton point is assigned a different ID number when stored in the file used for model training. Since the video is recorded at a frame rate of 25 FPS, the number of captured frames is calculated as 25 x video recording time in second. As a result, the total number of frames collected from the dataset of 16 action classes is 19,345 frames, as shown in Table III.

2) LSTMs for Patient Activity Recognition:

Long Short-Term Memory [25] is a Recurrent Neural Network (RNN) that has been increasingly used in the field of deep learning and human action recognition. LSTM, as opposed to standard feedforward neural networks, includes feedback connections. A recurrent neural network of this type can analyze not just individual data points (such as photographs), but also whole data sequences (such as speech or video). For instance, LSTM may be used for handwriting recognition, speech recognition, and anomaly detection in network traffic or intrusion detection systems. A typical LSTM unit comprises of a cell, an input port, an output port, and a forget port (shown in Fig. 5). The cell stores values for an indefinite amount of time, and the three gates control the flow of information into and out of the cell.

LSTM networks are well suited for classification, processing, and prediction based on time series data because they can handle indeterminate delays between significant events in time series. LSTM was developed to solve the vanishing gradient problem that can be encountered when training traditional RNNs. The benefit of LSTM over standard RNNs, Hidden Markov models, and other sequential learning approaches is its low sensitivity across a particular length range. RNNs can, in theory, follow any long-term relationships in input sequences. The problem with RNNs, however, is computational nature: when training an RNN using backpropagation, the backpropagation gradients can be degraded (i.e. tend to move towards zero) or "explode" towards infinity. Because LSTM units allow gradients to remain constant, RNNs utilizing LSTM units can partially alleviate the gradient degradation problem. However, these LSTMs can still suffer from gradient "explosion" problems.

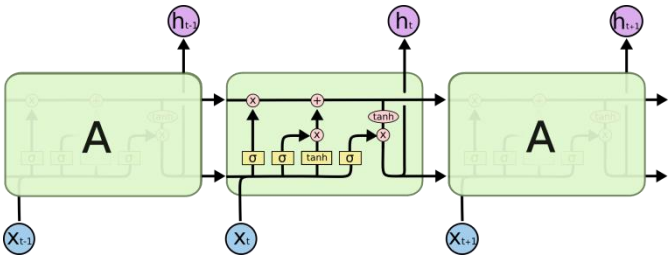


Fig. 5. Illustration of an LSTM Cell and Architecture. The Repeater Module in an LSTM Contains Four Interaction Layers.

In this work, a multi-layer LSTM with four layers have been implemented for patient activity recognition. Each layer has 50 units and is followed by a dropout layer designed to decrease the model's overfitting to the training data. Finally, a dense fully connected layer with 27 units is utilized to interpret the features retrieved by the LSTM hidden layer before making predictions with a final output layer with softmax function. The efficient Adam version of stochastic gradient descent will be utilized to optimize the network, and the categorical cross entropy loss function will be used because we are learning a multi-class classification problem.

D. Raspberry Pi and Camera Raspberry Pi NoIR

The Raspberry Pi [23] is a tiny computer developed by the Raspberry Pi Foundation in collaboration with Broadcom in the United Kingdom. The original Raspberry Pi project's goal was to promote basic computer literacy education in schools and developing countries. This device, however, became unexpectedly popular and was marketed for the purpose of building robots. Because of its inexpensive cost and open design, it is frequently utilized in various sectors, including weather monitoring. After the second version was released, the Raspberry Pi Foundation produced a brand-new gadget called the Raspberry Pi Trading. Raspberry Pi 4 Model B was released in June 2019 [33] with a 1.5 GHz quad-core ARM Cortex-A72 processor, 802.11ac Wi-Fi, Bluetooth 5, gigabit Ethernet (unlimited throughput), two USB 2.0 ports, two USB 3.0 ports, 2-8 GB RAM, and dual monitor support via a pair of

micro HDMI ports for up to 4K resolution. When used in conjunction with an appropriate power supply, the Raspberry Pi 4 is also powered via a USB-C port, allowing additional power to be provided to downstream peripherals (Fig. 6).

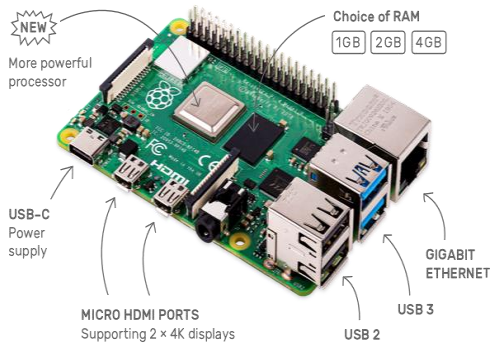


Fig. 6. Raspberry Pi 4 Model B.

The Raspberry Pi NoIR V2 IMX219 Camera (Fig. 7) is the latest version of the Camera Module for Raspberry Pi that uses the 8-megapixel IMX219 image sensor from Sony instead of the old OV5647 sensor. With the 8-megapixel IMX219 sensor from Sony, the Camera Module for Raspberry Pi has achieved a remarkable upgrade in both image and video quality as well as durability. Raspberry Pi NoIR V2 IMX219 8 MP camera can be used with Raspberry Pi to take photos and videos in low light conditions with HD 1080p30, 720p60, or VGA90 quality. It's also very simple, as we only need to connect the Raspberry Pi's Camera port and config to run the program. The Raspberry Pi NoIR V2 IMX219 8 MP camera is controllable via MMAL and V4L APIs, there are many libraries developed by the Raspberry Pi community on Python that make learning and using it much easier.

Fig. 7. Camera Raspberry Pi NoIR V2.

III. Experimental Results and Discussion

A. Experimental Settings and Evaluation Metric

All experiments were conducted on the laptop Asus TUF Gaming FX506LH i5 10300H, 8 GB RAM, NVIDIA GeForce GTX 1650 4GB, with the Ubuntu operating system. The real-time recognition system's algorithm is written in Python and implemented on the Linux operating system using the open-source libraries Keras and OpenCV (Raspberry Pi OS). Accuracy is a metric used to evaluate how well classification models perform and is calculated using the formula (1).

$$Accuracy = \frac{\text{Number of samples correctly classified}}{\text{Total number of samples}} \quad (1)$$

B. Model Training and Evaluation

The dataset of 16 action classes was split into two subsets that were used to train and evaluate the model at an 80:20 ratio, as was done by Luhach et al. [34]. Thus, the dataset has 15,476 frames for training and the remaining 3,869 frames for model evaluation. The system uses the data converted from video frames to coordinates (x, y, z) and the visibility value of the skeleton point features to train the model when using MediaPipe Pose. All this data is stored in CSV format file. Dataset for training and evaluating models are described in Table III.

TABLE III. Dataset Information for Model Training and Evaluation

Number of samples	Training	Testing	Timesteps	Dimension
19,345 frames	19,345 × 0.8 = 15,476 frames	19,345 × 0.2 = 3,869 frames	20	132 (33 points x 4 values) × Time-steps × number of frames

Since elaborate hyperparameter optimization methods like grid search were judged too time-consuming for the scope of this study, the various parameter-settings for the training process were developed via trial and error. As a result, many parameter choices have been attempted and tested by repeatedly running the model, and values for the hyperparameters that are thought to be near to an equilibrium between time-efficiency and performance have been chosen. In the selection process, several settings that reduced the difference between high and low values were tested, which is not unlike to how many root finding techniques in mathematics operate.

Different batch sizes (number of samples per gradient update) were examined, and 32 was found to be an appropriate value in terms of both effectiveness and performance. It was decided to iterate through the full dataset 50 times because epoch sizes greater than 50 produced negligible to no improvements. The mean squared error is employed for loss function. Adam produced the best results of the several optimizers available and was hence chosen over stochastic gradient descent. Finally, in the LSTM network model, the time-steps K are the most critical parameters affecting model performance. The time-steps are how many lagged variables the model receives as input to forecast the following step. Therefore, various number of time-steps are examined to determine how they impact the model's performance. The time-steps chosen to be tested are 5, 10, 15, and 20. The model is trained and evaluated five times for each of these time-steps in order to gather sufficient data to compare their relative performance. The performance of the model using different time-steps is illustrated in Table IV.

TABLE IV. The Overall Performance of the Model with Different Time-Steps.

K	5	10	15	20
Accuracy	92.26%	95.63%	96.44%	96.84%
Loss	0.1838	0.1047	0.0914	0.0854

Table IV shows that as the number of time-steps *K* is increased, the model's performance improves (accuracy increases and loss lowers), but training time increases. For instance, a model with 10 time-steps segments performs better than one with 5 time-steps (95.63 percent vs 92.26 percent). This result demonstrates that the model will learn more information from earlier frames if many time-steps *K* are used. As a result, the resulting features of the videos will be more robust and high-abstract, improving the model's classification precision. However, when *K* is increased to 20, the model performance exhibits evidence of saturation at 96.44% as opposed to 96.44% with *K* = 15. Therefore, we set *K* equal to 20 for the model to achieve the best classification performance while keeping training and evaluation time to a minimum. Fig. 8 depicts the curve reflecting the model's accuracy and loss after 50 iterations.

After training, the resulting model size is only 1.2 MB, making it appropriate for installation on Raspberry Pi devices with limited memory configuration. According to the article [22], the FPS of BlazePose Full is 102, while that of BlazePose Lite is 312, making it ideal for developing real-time applications. The model training process is quite fast, averaging about 5–10 seconds for an iteration with 16 action classes, because the video data has been converted to a text file in CSV format, so it doesn't take a lot of hardware resources. The model is trained in 50 iterations taking from 5 to 10 minutes. The resulting model has achieved an accuracy of 96.84% on our dataset. After the real-time action recognition model's training and evaluation on the Rapberry Pi 4 system produced good and consistent results, the system was installed and tested to send alert messages and emails if the camera detects unusual patient health-related behaviors (shown as shown in Fig. 9 and Fig. 10). Fig. 11 depicts our system's successful detection of six real-time patient actions: hand clap, sit down, stomachache, backache, and fall down. Our next step is to collect more data from a variety of patient actions and then to investigate mobile devices with better hardware configurations, such as the Jetson Nano Developer Kit [35] and CNN models that are efficient, accurate, and suitable for the latest mobile devices.

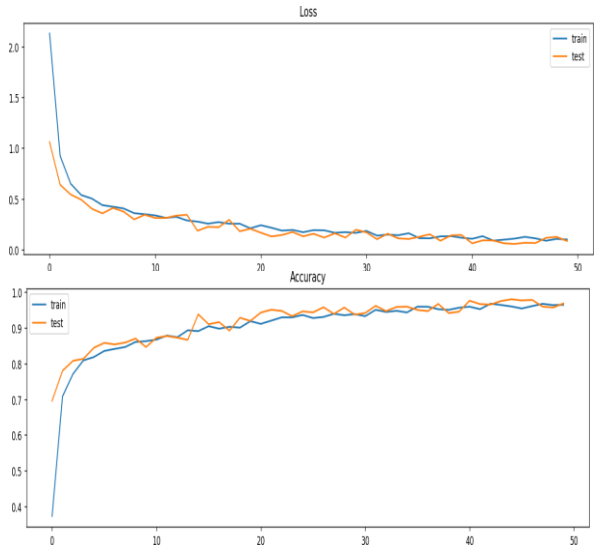


Fig. 8. Accuracy and Loss of Training Process with 50 Epoches.

Fig. 9. Patient Alert Emails are sent to the Users.

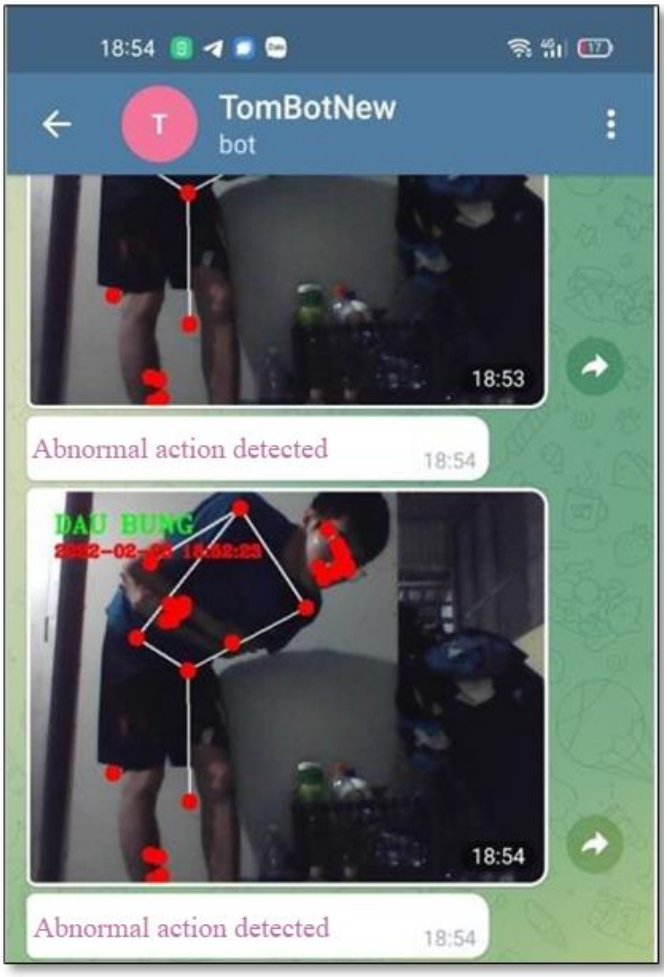


Fig. 10. Alert Messages and Images are sent to the Users via the Telegram Application.

Fig. 11. Real Time Patient Action Identification Results on Raspberry pi 4, Camera Module Noir System.

IV. Conclusion and Future Research Work

This study has proposed a novel system with basic functions of a smart surveillance camera, supporting remote patient monitoring. A model for remote skeletal patient activity detection was developed using MediaPipe Pose, an LSTM network, and a Raspberry Pi 4. The numerical results show that our proposed model performed well in classification, with an accuracy of 96.84% on a dataset of 16 activities gathered and constructed by ourselves. In addition, because the MediaPipe Pose library and the LSTM network are used for recognition, the recognition model size is small, and the network training parameters are few, making it appropriate for deployment on mobile devices with limited hardware, such as the Raspberry Pi 4. Therefore, our method offers numerous benefits in terms of real-time patient action recognition, low cost, simple installation, and practical implementation. A dataset of 541 video files of patients' actions in indoor was built to evaluate our method. Although the amount of data is little and there isn't much actual patient data, this provides the foundation for future larger, better-quality data sets that will help the research community better understand patient activities.

References

[1] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," Image Vis. Comput., vol. 60, pp. 4–21, 2017, doi: 10.1016/j.imavis.2017.01.010.

[2] Y. Bengio, "Deep Learning of Representations: Looking Forward," ArXiv, vol. abs/1305.0, 2013.

[3] C. Szegedy et al., "Going deeper with convolutions," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June, no. January 2017, pp. 1–9, 2015, doi: 10.1109/CVPR.2015.7298594.

[4] J. Gutiérrez, V. Rodríguez, and S. Martin, "Comprehensive review of vision-based fall detection systems," Sensors (Switzerland), vol. 21, no. 3, pp. 1–50, 2021, doi: 10.3390/s21030947.

[5] L. P. Malasinghe, N. Ramzan, and K. Dahal, "Remote patient monitoring: a comprehensive study," J. Ambient Intell. Humaniz.

Comput., vol. 10, no. 1, pp. 57–76, 2019, doi: 10.1007/s12652-017-0598-x.

[6] P. Vallabh and R. Malekian, “Fall detection monitoring systems: a comprehensive review,” J. Ambient Intell. Humaniz. Comput., vol. 9, no. 6, pp. 1809–1833, 2018, doi: 10.1007/s12652-017-0592-3.

[7] R. Rucco et al., “Type and location of wearable sensors for monitoring falls during static and dynamic tasks in healthy elderly: A review,” Sensors (Switzerland), vol. 18, no. 5, 2018, doi: 10.3390/s18051613.

[8] Z. Liu, Y. Cao, L. Cui, J. Song, and G. Zhao, “A Benchmark Database and Baseline Evaluation for Fall Detection Based on Wearable Sensors for the Internet of Medical Things Platform,” IEEE Access, vol. 6, pp. 51286–51296, 2018, doi: 10.1109/ACCESS.2018.2869833.

[9] S. Cheng, L. Thomas, J. Cook, and M. Pecht, “A Radio Frequency Sensor System for Prognostics and Health Management,” 2009, doi: 10.1115/DETC2009-87723.

[10] M. Jang, S. Kang, and S. Lee, “Monitoring Person on Bed Using Millimeter-Wave Radar Sensor,” in 2022 IEEE Radar Conference (RadarConf22), 2022, pp. 1–4, doi: 10.1109/RadarConf2248738.2022.9764251.

[11] Q. Li, W. Cai, X. Wang, Y. Zhou, D. Feng, and M. Chen, “Medical image classification with convolutional neural network,” 2014 13th Int. Conf. Control Autom. Robot. \& Vis., pp. 844–848, 2014.

[12] J. Gu et al., “Recent advances in convolutional neural networks,” Pattern Recognit., vol. 77, no. June 2016, pp. 354–377, 2018, doi: 10.1016/j.patcog.2017.10.013.

[13] M. Z. Uddin and M. Hassan, “Activity Recognition for Cognitive Assistance Using Body Sensors Data and Deep Convolutional Neural Network,” IEEE Sens. J., vol. 19, pp. 8413–8419, 2019.

[14] A. D. Ignatov, “Real-time human activity recognition from accelerometer data using Convolutional Neural Networks,” Appl. Soft Comput., vol. 62, pp. 915–922, 2018.

[15] C. Avilés-Cruz, A. Ferreyra-Ramírez, A. Zúñiga-López, and J. Villegas-Cortéz, “Coarse-fine convolutional deep-learning strategy for human activity recognition,” Sensors (Switzerland), vol. 19, no. 7, 2019, doi: 10.3390/s19071556.

[16] W.-H. Chen, C. Baca, and C.-H. Tou, “LSTM-RNNs combined with scene information for human activity recognition,” 2017 IEEE 19th Int. Conf. e-Health Networking, Appl. Serv., pp. 1–6, 2017.

[17] H. Li and M. Trocan, “Personal Health Indicators by Deep Learning of Smart Phone Sensor Data,” 2017 3rd IEEE Int. Conf. Cybern., pp. 1–5, 2017.

[18] T. Thanh-Nghi, D., Thanh-Hien-Triet, N., Truong-An, “Smart camera system for remote patient activity monitoring,” in the 14th National Scientific Conference on Research and Application of Information Technology - Fair’ 2021, Natural Science and Technology Publishing House, 2021, pp. 110–117.

[19] H. Ramirez, S. A. Velastin, I. Meza, E. Fabregas, D. Makris, and G. Farias, “Fall Detection and Activity Recognition Using Human Skeleton Features,” IEEE Access, vol. 9, pp. 33532–33542, 2021, doi: 10.1109/ACCESS.2021.3061626.

[20] J.-C. Chiang et al., “Posture Monitoring for Health Care of Bedridden Elderly Patients Using 3D Human Skeleton Analysis via Machine Learning Approach,” Appl. Sci., vol. 12, no. 6, p. 3087, 2022, doi: 10.3390/app12063087.

[21] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, “Skeleton-based human activity recognition using ConvLSTM and guided feature learning,” Soft Comput., vol. 26, no. 2, pp. 877–890, 2022, doi: 10.1007/s00500-021-06238-7.

[22] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, “BlazePose: On-device Real-time Body Pose tracking,” CoRR, vol. abs/2006.1, 2020, [Online]. Available: https://arxiv.org/abs/2006.10204.

[23] W. Gay, Raspberry Pi Hardware Reference. Apress Berkeley, CA, 2014.

[24] C. Lugaresi et al., “MediaPipe: A Framework for Building Perception Pipelines,” 2019, [Online]. Available: http://arxiv.org/abs/1906.08172.

[25] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[26] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, “ActivityNet: A large-scale video benchmark for human activity understanding,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 07-12-June. pp. 961–970, 2015, doi: 10.1109/CVPR.2015.7298698.

[27] W. Kay et al., “The Kinetics Human Action Video Dataset,” ArXiv, 2017, [Online]. Available: http://arxiv.org/abs/1705.06950.

[28] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild,” no. December 2012, 2012, [Online]. Available: http://arxiv.org/abs/1212.0402.

[29] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “HMDB: A large video database for human motion recognition,” Proc. IEEE Int. Conf. Comput. Vis., pp. 2556–2563, 2011, doi: 10.1109/ICCV.2011.6126543.

[30] Y. Yoshikawa, J. Lin, and A. Takeuchi, “STAIR Actions: A Video Dataset of Everyday Home Actions.” 2018, [Online]. Available: http://arxiv.org/abs/1804.04326.

[31] S. Gaglio, G. Lo Re, and M. Morana, “Human Activity Recognition Process Using 3-D Posture Data,” IEEE Trans. Human-Machine Syst., vol. 45, no. 5, pp. 586–597, 2015, doi: 10.1109/THMS.2014.2377111.

[32] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, “NTU RGB+D: A large scale dataset for 3D human activity analysis,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-Decem. pp. 1010–1019, 2016, doi: 10.1109/CVPR.2016.115.

[33] E. Upton, “Raspberry Pi 4 on sale now from $35.” Raspberry Pi Foundation, 2019.

[34] A. Luhach, D. Jat, K. Ghazali, X.-Z. Gao, and P. Lingras, Advanced Informatics for Computing Research: Third International Conference, ICAICR 2019, Shimla, India, 15–16 June 2019; Revised Selected Papers; Springer Nature: Berlin/Heidelberg, Germany, 2019; Volume 1075. 35. 2019.

[35] NVIDIA, “Jetson Nano Developer Kit User Guide.” 2021, [Online]. Available: https://developer.download.nvidia.com/ embedded/L4T/r32-3-1_Release_v1.0/Jetson_Nano_ Developer_Kit_User_Guide.pdf.

A Covid-19 Positive Case Prediction and People Movement Restriction Classification

I Made Artha Agastya
Informatics
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia

***Abstract*—The world experienced a pandemic that changed people's daily life due to Coronavirus Disease 2019 (covid-19). In Jakarta, the covid-19 cases were discovered on March 18, 2020, and the case increased uncontrollably until the government conducted a movement restriction called pembatasan sosial berskala besar (PSBB). The effectivity of movement restriction was not evaluated in detail. Therefore, we investigated the covid-19 cases in the PSBB period to understand the contribution of movement restriction. Moreover, a prediction model is proposed to computerize the decision of movement restriction. The models are divided into regression and classification models. The regression model is developed to forecast the number of infected cases. At the same time, the classification model is used to identify the best movement restriction type. We utilize data transformation named Principal Component Analysis (PCA) to reduce the number of features. In our case, the best regression method is Multiple Linear Regression (MLP). Then, the best classification method is the Support Vector Machine (SVM). The MLP results are 148.38, 37036.37, and 0.250336 for Mean Absolute Error (MAE), Mean Square Error (MSE), and R^2, respectively. In contrast, the SVM achieved an accuracy of 84.81%. Moreover, the prediction system on the website were successfully deployed.**

Keywords*—Covid-19; movement restriction; machine learning; positive case; infected prediction

I. Introduction

The Coronavirus Disease 2019 (COVID-19) was first detected in China at the end of 2019. The virus causes breathing symptoms such as fever, coughing, pneumonia, and diarrhea in patients [1]. The World Health Organization (WHO) received a report from the Chinese government on December 31, 2019, regarding this disease with reports on the case of pneumonia or wet lungs in Wuhan, in Hubei Province, China. A week later, the Chinese government confirmed that Covid-19 had been identified as the cause of pneumonia [2]. Then the media reported that many new cases were recorded in other countries because international travel and trade were operating as usual.

As the center of government and trade in Indonesia, Jakarta has the highest population density compared to other places [3]. Consequently, an increased number of positive confirmed Covid-19 in Jakarta. Based on data on March 18, 2020, there were 171 cases. In Jakarta, the number of positive cases of Covid 19 drastically expanded since there were ten times increasing positive cases on April 9, 2020, compared to March 18, 2020. Jakarta suffered 1776 cases at that time. While the average increase per day for the period March 18 until April 9, 2020, is 70 cases. However, at that time, the government had not taken any corrective actions. In contrast, Malaysia held a national scale MCO (Movement Control Order), which began on March 18, 2020 [4]. After three weeks without corrective action, then on April 10, 2020, the DKI Jakarta Regional Government imposed a pembatasan sosial berskala besar (PSBB) consisting of two categories, namely a strict PSBB and a transition PSBB. The Strict PSBB was valid from April 10 - June 4, 2020, and September 14 - October 11, 2020. And the Transition PSBB took place from June 5 - September 10, 2020, and October 12 - October 25, 2020.

Based on [5], [6], the positive confirmation trend per day when Strict PSBB is carried out tends to be stable when compared to No-PSBB. This shows the effect of restrictions on community movements. To reduce the economic burden, the government relaxed community activities by imposing Transition PSBB. However, there is a significant increase in the number of positive cases per day. Comparing the Transition PSBB and No-PSBB, the condition of the Transition PSBB is more worrying as the rate, and cumulative cases are much higher. A comprehensive data collection can be an excellent choice to gain more insight. Moreover, an artificial intelligence system can help the government to decide the type of movement restriction. Therefore, the research question can be determined as follow:

1) What factors or features affect the number of daily positive cases?

2) How to predict the number of daily positive cases?

3) How to determine the type of movement restriction in certain conditions?

This paper is organized into five sections which are the introduction, related works, methods, result and discussion, and conclusion and future work. The introduction section consists of a brief explanation of the problem and the purpose of investigation. The related work section contains the summary of other researcher works that solved a covid-related prediction problem. The methods section shows the data acquisition, preparation, exploration, transformation, and frameworks. After that, the result and discussion section report the experimental findings. The conclusion and future work section contain a final comment on the current work and the potential research in the future.

II. Related Work

The infectious disease outbreaks such as food contamination, severe acute respirations syndrome (SARS), dengue, malaria, and influenza [7]–[9] have been analyzed using the artificial intelligence (AI) approach. The most common research is to forecast the number of infected individuals using the machine learning method. As the data is mostly time series, the model can be following a linear or nonlinear model [10]–[12]. The AI system can be used to prevent the outbreak become uncontrollable as the estimated number of diseased people is calculated. The government can prepare the medical facilities and their staff to match the number of patients. The anticipation of the worst-case scenario makes us more ready to confront the outbreak.

The machine learning-based model [13] has been deployed to predict the positive case of covid-19. The result is satisfying as the R^2 of 0.9998, 0.9996, and 0.9999 for Gompertz, Logistic, and Artificial Neural Networks models, respectively. Based on the results, trends can be forecasted and extended until the termination of the pandemic in Mexico. However, their models were only able to predict new COVID-19-positive cases. Velásquez and Lara [14] utilized a reduced-space Gaussian process regression model to forecast 82 days of positive, dead, and recover cases in the USA. Compared to the actual values, it is discovered that the model can generate the expected case value with a significant correlation coefficient. Buckingham-Jeffer [15] performs prediction using stochastic SIR (susceptible infectious removed) and SEIR (susceptible-exposed infectious removed) models. Both models can immediately forecast the number of positive cases using maximum likelihood inference in a time frame period.

A typical time-series experiment uses the Autoregressive Integrated Moving Average (ARIMA) model to produce a prediction model to estimate the amount of COVID-19 confirmed cases, fatalities, and recoveries [16]. If the current trend in Pakistan continues, the number of actual instances might triple by May. The investigation result concludes that the statistics are accurate and that the trends will continue to rise in the next month. Fortunately, the ARIMA model cannot comprehend the provided historical data pattern, so their findings are questionable.

A comparison of ARIMA with machine learning methods has been conducted by Kamarudin et al. [17] for positive, dead, and recovery cases in Malaysia. They compare support vector regression (SVR), Gaussian process (GP), linear regression (LR), neural network (NN), and ARIMA. They found that the NN outperforms most of the methods in positive and recovery case prediction. The ARIMA can predict the number of dead cases accurately. The most unreliable model is LR because it overestimated the case number so much. Meanwhile, the GP and SVR have shown a promising result as the Root Mean Square Error (RMSE) value is around the ARIMA value.

The previous literature shows the machine learning method can forecast the outbreak damage level by predicting the number of infected people. In this study, the number of positive cases is not only expected but the type of movement restriction can be predicted using the machine learning method. The move restriction has been proved by many countries can reduce the number of positive and dead cases [4], [18]–[23]. The comparison of Multiple Linear Regression (MLR), Support Vector Regression (SVR), Random Forest Regressor, and Decision Tree Regressor are conducted to know the best regression method [24]–[26]. In contrast, the classification methods are implemented to determine the movement restriction status. The Logistic Regression, Support Vector Machine (SVM), Decision Tree, Naïve Bayes, KNN, Adaboost, and XGboost methods are compared to know the most successful approach to predict the correct movement restriction. Moreover, the Particle Component Analysis (PCA) and grid search algorithm are implemented to understand the features reduction and hyperparameter optimization effect in accuracy, Mean Absolute Error (MAE), and Mean Square Error (MSE) [27]–[29].

The number of positive cases can be predicted by using a regression model, and the number of expected cases is combined with factual information (e.g., number of public transport passengers, wind speed, and congestion level) at that time to predict the type of the restriction. The number of positive cases needs to be predicted because the data of positive cases is unknown at that specific time. Therefore, a prototype of the website application is developed and deployed to show the proposed system in a real environment.

III. Method

We collect all data such as COVID-19 patient data, weather, and climate data as well as air transportation user data. Then the data is cleaned using imputation method to recover the missing data and scaling method if the data is not on the same scale. Then the clean data is processed with statistical and visual analysis. Only then after getting a deeper understanding, the data can be added or reduced depending on the findings with visualized data. After the data is ready, the data can be trained with a regression algorithm to obtain predictions of the number of patients with Covid 19 in daily basis. The results of the prediction are entered into the classification system and the type of restrictions can be determined. The mean absolute error (MAE), the mean square error (MSE), and accuracy are used as measuring instrument.

A. Data Accuisation

Based on the problems formulation, we collect data that might affect the number of positive cases. To get complete COVID-19 data in Jakarta we access the following website: https://corona.jakarta.go.id. The display form of the website above is shown in Fig. 1. We can see various information related to Covid-19 in Jakarta such as the number of specimens being tested, the number of people who are vaccinated, the amount that recovered, etc.

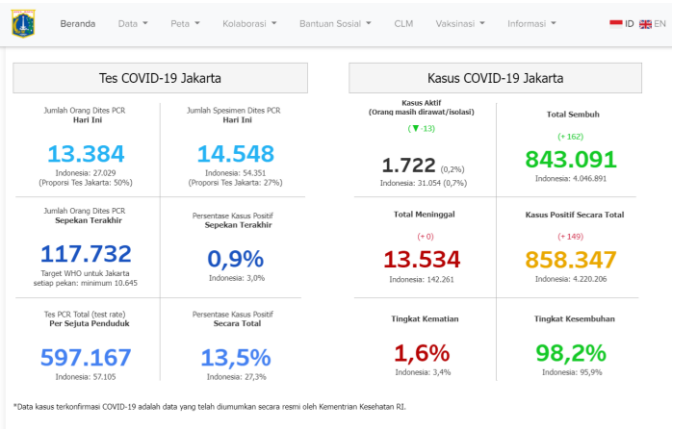


Fig. 1. The Main Menu of https://corona.jakarta.go.id.

Because we cannot access the data from the website, we traced the data source and found that the dashboard was made using the tableau application. Then we found the tableau address as follows:

https://public.tableau.com/app/profile/jsc.data/viz/dashboardcovid-19jakarta_15837354399300/dashboard22

From the website, we can download data in the form of a PDF File, as shown in Fig. 2. Then, from the PDF data, we change it to Excel File with the help of the PDF Editor application (Nitro Pro).

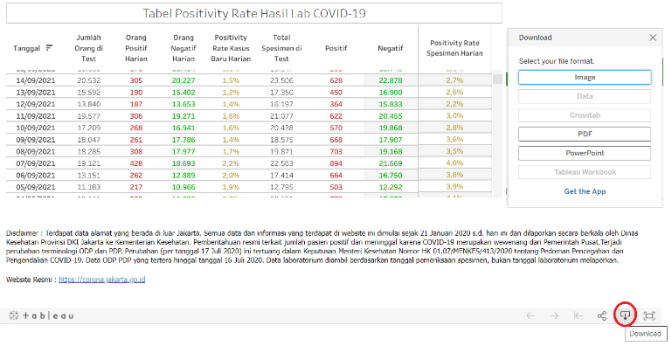


Fig. 2. Covid-19 Data in Jakarta at Tableau Site.

Then we searched the movement data of Jakarta people by looking at the number of users of public transportation modes. For train user data, we take from the following site: https://www.bps.go.id/indicator/17/72/2/sum-penumpang-kereta-api.html. From that website, we can download the data in an excel file, as shown in Fig. 3.

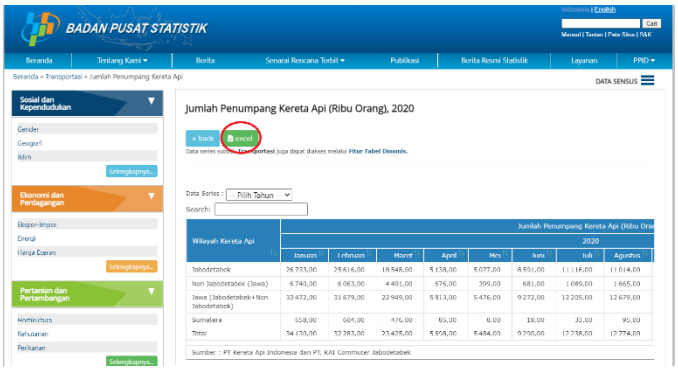


Fig. 3. The Data of Train Users in Jakarta.

Then we took the data of users of Transjakarta, Mass Rapid Transport (MRT), and LRT (Integrated Rail Cross) via https://data.jakarta.go.id/dataset. In addition to the three-transportation data, we can also obtain other data, such as data on the number of residents moving to Jakarta, the number of family planning participants, and others, as shown in Fig. 4.

Fig. 4. Jakarta Open Data Site.

We can perform a search, as shown in Fig. 4, and then select the data we want to access. For MRT data, we can download the data by clicking Download Data, as shown in Fig. 5. Then, we get a comma-separated value (CSV) from it. For the LRT and Transjakarta data, we used the same method as previously described to obtain the data. The addresses of the datasets are as follows:

https://data.jakarta.go.id/dataset/data-penumpang-transjakarta-2020/

https://data.jakarta.go.id/dataset/data-penumpang-mrt-di-provinsi-dki-jakarta

https://data.jakarta.go.id/dataset/data-penumpang-lrt-di-provinsi-dki-jakarta

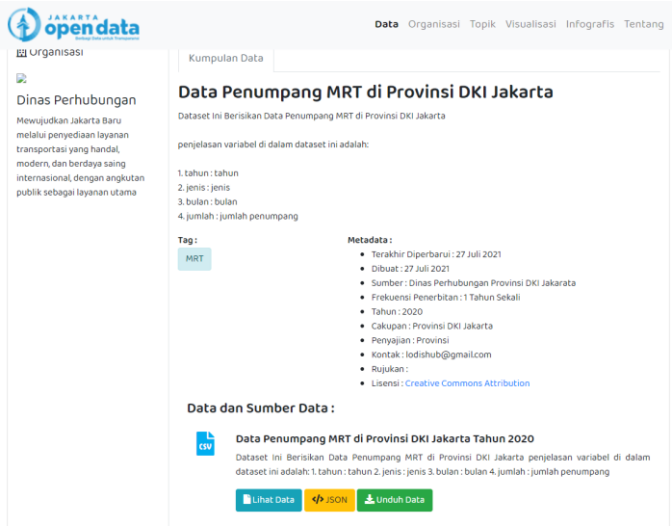


Fig. 5. Download Section in MRT, Jakarta Open Data.

For data on weather conditions and air quality in Jakarta, we take from two sources, namely:

1) https://dataonline.bmkg.go.id/data_iklim
2) https://aqicn.org/data-platform/covid19/

For Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) online data, we need to create an account first, and we can only download it for a period of 30 days, as shown in Fig. 6. While on the Air Quality Open Data Platform we can download air condition data around the world from 2015 to 2021 as shown in Fig. 7. From BMKG we can obtain data on rainfall, duration of exposure, wind speed, humidity, etc. Meanwhile, on the Air Quality Open Data Platform, we collected PM10 and PM2.5 data, which are international standard air condition indexes.

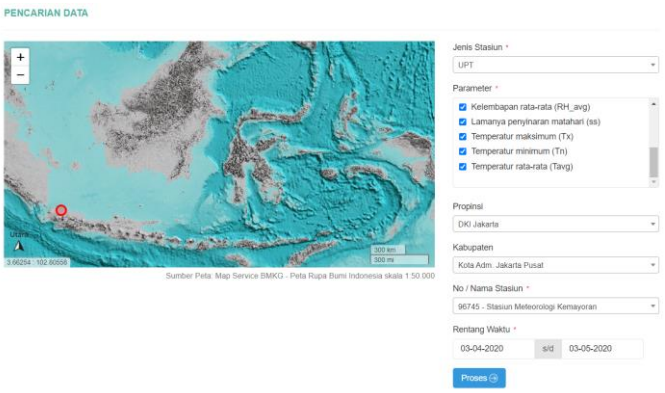


Fig. 6. Badan Meteorologi, Klimatologi, Dan Geofisika (BMKG) Data.

For congestion data, we take from https://www.tomtom.com/en_gb/traffic-index/jakarta-traffic/. We can only take data every month as an indicator of congestion, as shown in Fig. 8. The index is between 0 – 1, where 0 is the value without congestion and 1 is a great traffic jam.

For data on the status of people's movement restrictions, we take the following news:

https://news.detik.com/berita/d-5167032/timeline-psbb-jakarta-to-tarik-rem-darurat

From this news, we can determine the period of strict PSBB, transitional PSBB, and no PSBB. So that the data we need is complete but still not well organized and validated.

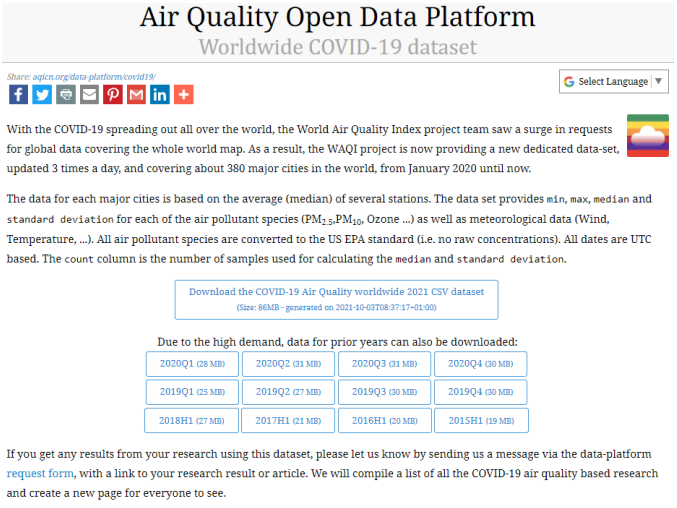


Fig. 7. Air Quality Open Data Platform

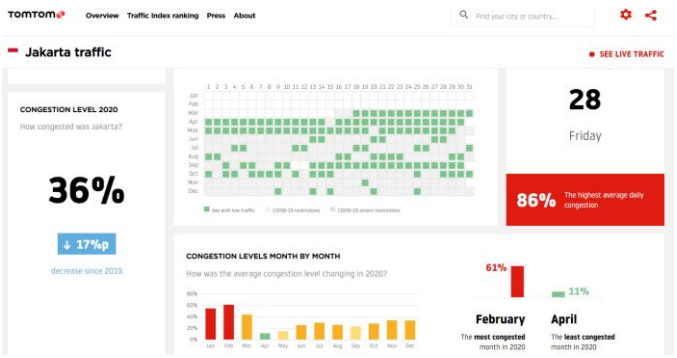


Fig. 8. Jakarta Congestion Data

B. Data Preparation

Once we get all the required data, we look at the data. The first criterion used is the availability of data during the PSBB period, which is March 10, 2020 – November 25, 2020. After investigating the dataset, the downloaded data meets these criteria. Then we discard the unnecessary information, such as the highest wind direction, maximum wind speed, maximum/minimum temperature, etc. We delete the data because we only used average values such as average wind speed and average temperature. In addition, we also discard data on the number of suspects, the number of negative covid-19, and the daily percentage of cases because this has been explained in the daily positive number and daily test number. After the needless data is removed, we then combine the data into a comprehensive table.

The overview of this research problems are the daily positive case prediction and the movement restriction type classification problems. We expect that the predicted positive case will be used as pseudo data to indicate the type of people's movement restrictions. Table I contains information

as follows, KRL Passenger, MRT Passenger, LRT Passenger, Bus Passenger, Congestion level, Rain drop rate, Day time, Humidity, Pm10, Pm2.5, Temperature, Wind, Restriction Status, Tested sample, and Positive Case.

TABLE I. COMBINATION OF ALL DATA

Date	KRL	..	Status	Tested Sample	Positive Case
10/03/2020	618266	..	NO_PSBB	52	2
11/03/2020	618266	..	NO_PSBB	48	26
12/03/2020	618266	..	NO_PSBB	56	10
13/03/2020	618266	..	NO_PSBB	30	7
14/03/2020	618266	..	NO_PSBB	44	16
15/03/2020	618266	..	NO_PSBB	41	2
..	..	..	..	..	..
25/11/2020	387400	..	NO_PSBB	15,381	1,124

C. Data Exploration

Descriptive statistical analysis can be seen in Table II below. From Table II, we can find out the average, maximum value, and minimum value of each attribute or feature. From Table II, there is nothing strange about these statistical values, so we can say that the combined data is valid.

TABLE II. STATISTICS OF ALL DATA

	KRL	MRT	LRT	Bus	..	Positive Case
count	261	261	261	261	..	261
mean	326717.6	14614.85	727.3257	251539.8	..	6080.831
std	117980.5	11273.04	461.2927	131444.5	..	4425.058
min	169233	1451	198	80396	..	30
25%	286366	11351	613	149401	..	2211
50%	337600	12991	632	285011	..	5295
75%	370533	17393	872	307195	..	10021
max	618266	46787	2024	559231	..	17871

Fig. 9 shows that the daily positive number changes along with the number of people being tested (blue) for COVID-19 per day. This indicates that the daily positive number (red) is strongly influenced by the number of people who test for Covid.

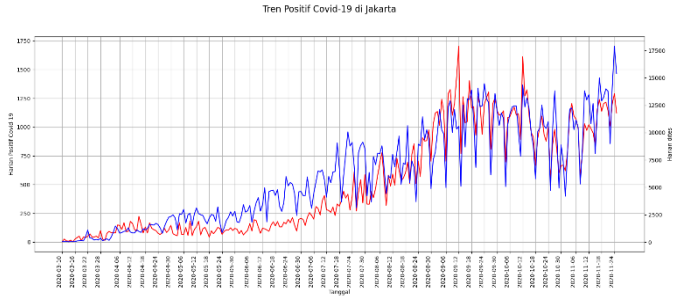


Fig. 9. The Positive Case vs the Sample Tested.

Fig. 10 shows a strong correlation between the number of daily tests and daily positives. There is a weak correlation between the number of Bus passengers and the daily test. There is also a weak correlation between congestion levels and daily tests.

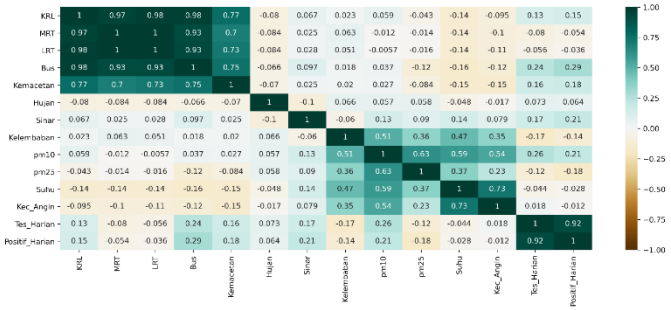


Fig. 10. The Correlation Matrix.

Based on Fig. 11 the distribution shows a positive correlation between the number of daily tests and the number of daily positives.

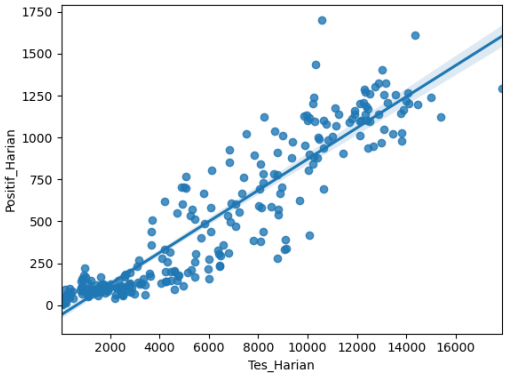


Fig. 11. Tested Sample vs. Positive Case.

Since we don't have daily data on congestion levels, we use monthly data, creating a scatter plot graph as shown in Fig. 12. There is a weak correlation between congestion with a positive total.

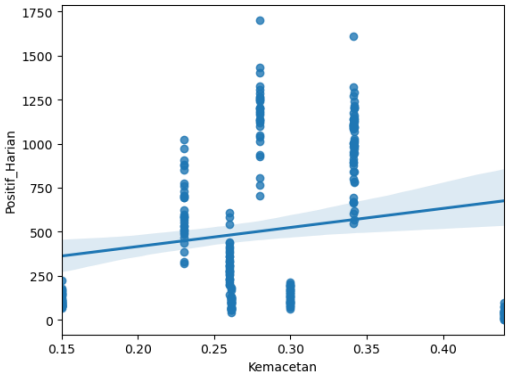


Fig. 12. Congestion vs. Positive Case.

Because data on the number of bus passengers per day is not available, we use the average per day of the number of passengers per month as shown in Fig. 13. We get a weak positive correlation between Bus vs. Total positive.

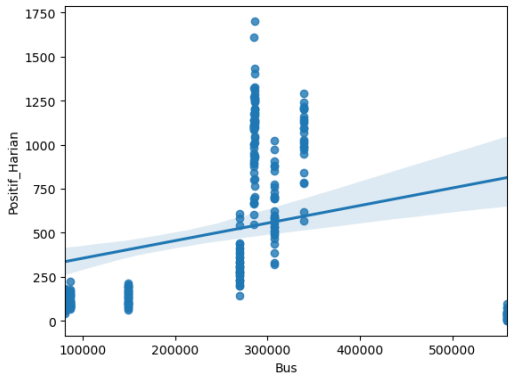


Fig. 13. Bus Passenger vs. Positive Case.

There is a very weak correlation between solar irradiance and daily positive, as shown in Fig. 14.

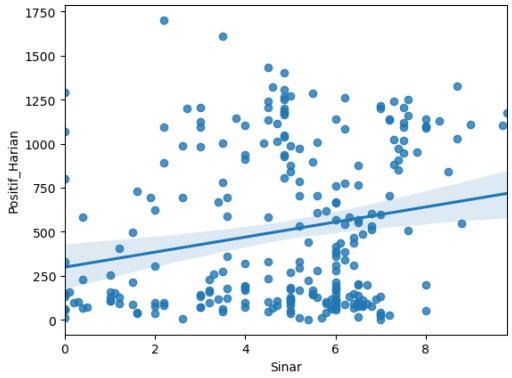


Fig. 14. Day Time vs. Positive Case.

There is a very weak negative correlation between wind speed and positive total, as shown in Fig. 15.

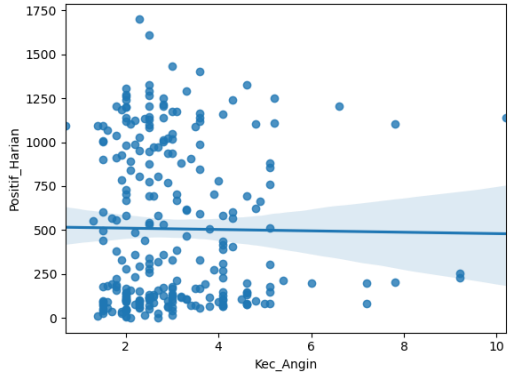


Fig. 15. Wind Speed vs. Positive Case.

Based on Fig. 16, there is a difference in the median value between no PSBB and strict PSBB. However, there was no significant difference between the three PSBB statuses in the number of daily positive cases.

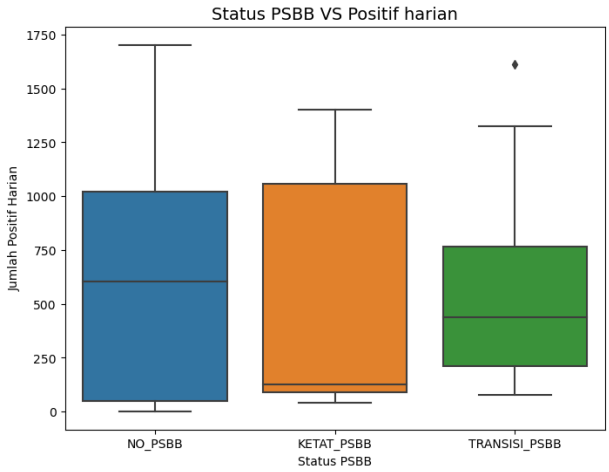
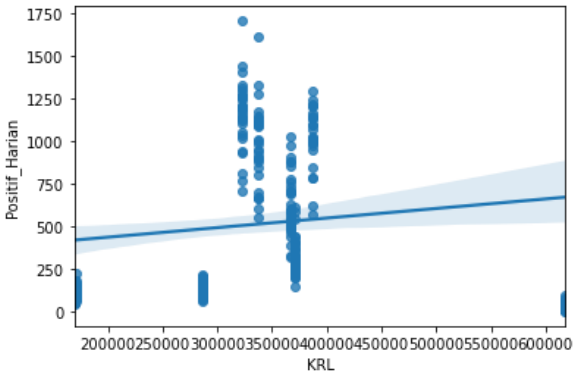


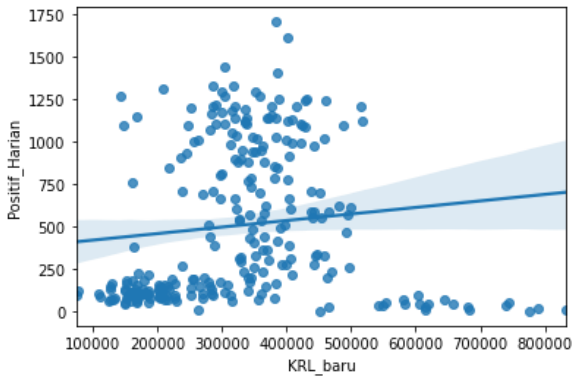
Fig. 16. Block Plot of People Restriction Status.

D. Data Transformation

Since we do not have daily data on the number of KRL passengers, we use the mean of total monthly passengers. However, giving the same value for one month is very unrealistic, so we need to transform the data to be more realistic by generating random data that meets the criteria for the mean of total passengers and standard deviation = 0.2 * mean, as shown in Fig.17.



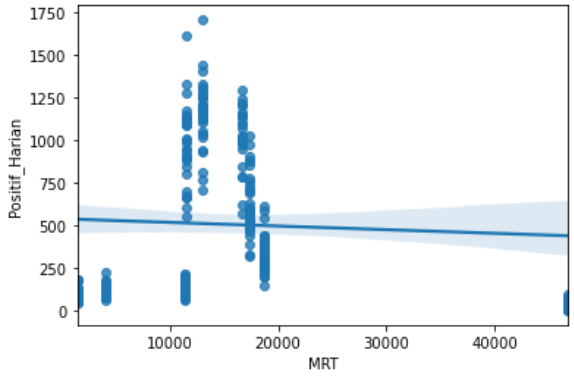
(a)



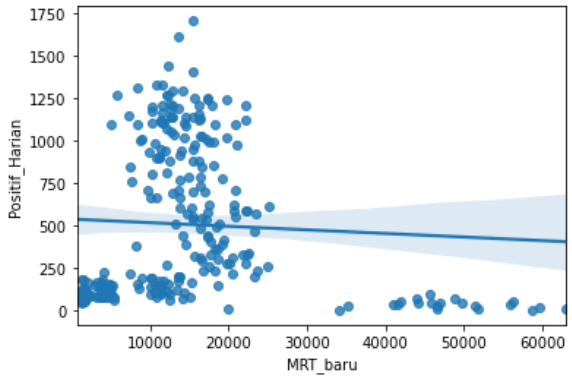
(b)

Fig. 17. KRL Passenger Data Transformation.

To be able use total monthly passengers, number of LRT, number of Bus passengers, and daily data of congestion in realistic way, data transformations must be done. The random data are generated using mean and standard deviation (it is calculated by 0.2 * mean) of the respective datasets. The transformations of the dataset have shown in Fig. 18, Fig. 19, Fig. 20, and Fig. 21 respectively.

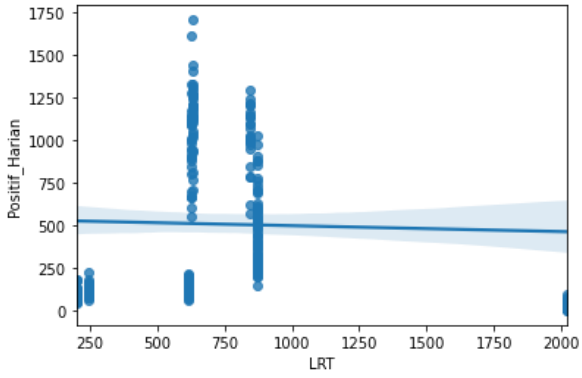


(a)

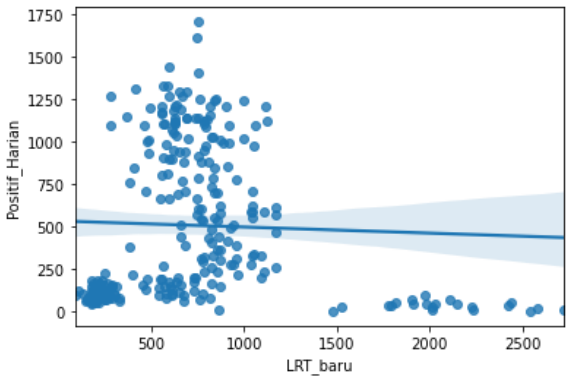


(b)

Fig. 18. MRT Passenger Data Transformation. (a) Before Transformation, (b) After Transformation.

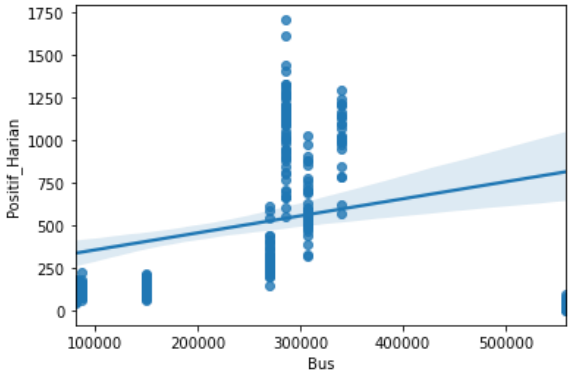


(a)

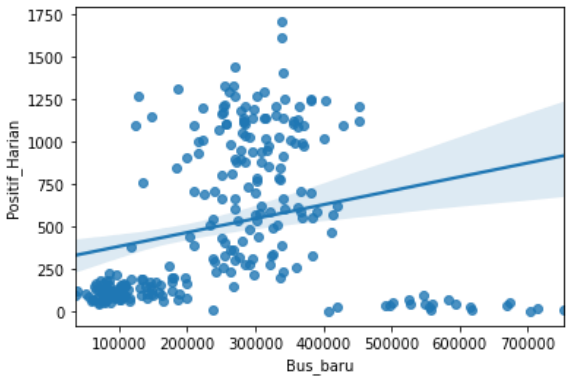


(b)

Fig. 19. LRT Passenger Data Transformation. (a) Before Transformation, (b) After Transformation.

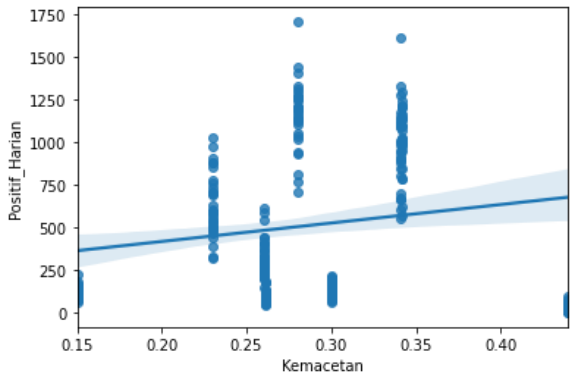


(a)

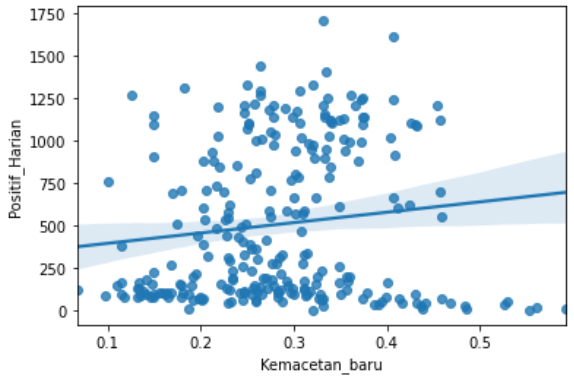


(b)

Fig. 20. Bus Passenger Data Transformation. (a) Before Transformation, (b) After Transformation.



(a)



(b)

Fig. 21. Congestion Data Transformation. (a) Before Transformation, (b) After Transformation.

To be able to see the correlation between PSBB status and other attributes, it is necessary to transform the data from PSBB status with one hot encoding. Then by using the Pearson correlation method, a correlation table is obtained as shown in Fig. 22.

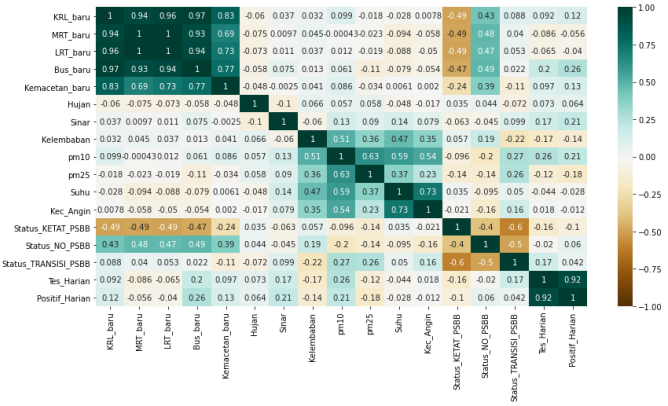


Fig. 22. Correlation Analysis after Data Transformation.

Based on Fig. 22, the daily test has the most significant correlation, which is 0.92. Then Bus_baru, which is the number of bus passengers per day, has a weak correlation of 0.26. Then the light, which is the length of irradiation in hours, has a correlation of 0.21. Then pm10, the air pollution index, has a weak correlation of 0.21. And humidity has a weak correlation of -0.14.

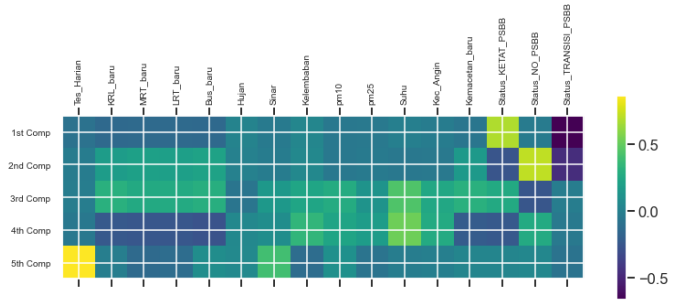


Fig. 23. The Contribution of Each Feature to PCA Value.

Based on Fig. 23, the strict PSBB status gave a positive contribution to the 1st component. In addition, the No PSBB status made a positive contribution to the 2nd component. The transitional PSBB status made a negative contribution to 1st component. Then the daily test made a very large contribution to the 5th component. The top five of Principle Component Analysis (PCA) value for each component can be found in Table III.

TABLE III. PCA VALUE

1st Comp	2nd Comp	3rd Comp	4th Comp	5th Comp
-0.30117	1.454034	0.571142	-0.52336	-0.3233
-0.18254	1.251233	0.288714	-0.26551	-0.23284
-0.22774	1.363212	0.258748	-0.54626	-0.30674
-0.33262	1.537455	0.586097	-0.69514	-0.45323
-0.33639	1.458526	0.629993	-0.52833	-0.28187

E. Framework

Based on the aim of this project is to create a system that can predict the number of people infected with Covid-19 and determine what restrictions should be made. Therefore, the PSBB status and the number of daily positive confirmations are labeled to be predicted. We created two models, namely the regression model and the classification model. The Regression Model is used to indicate the daily confirmed number of Covid-19 under certain conditions, and the classification model is used to determine what restrictions should be placed on the state of the daily positive count, as shown in Fig. 24.

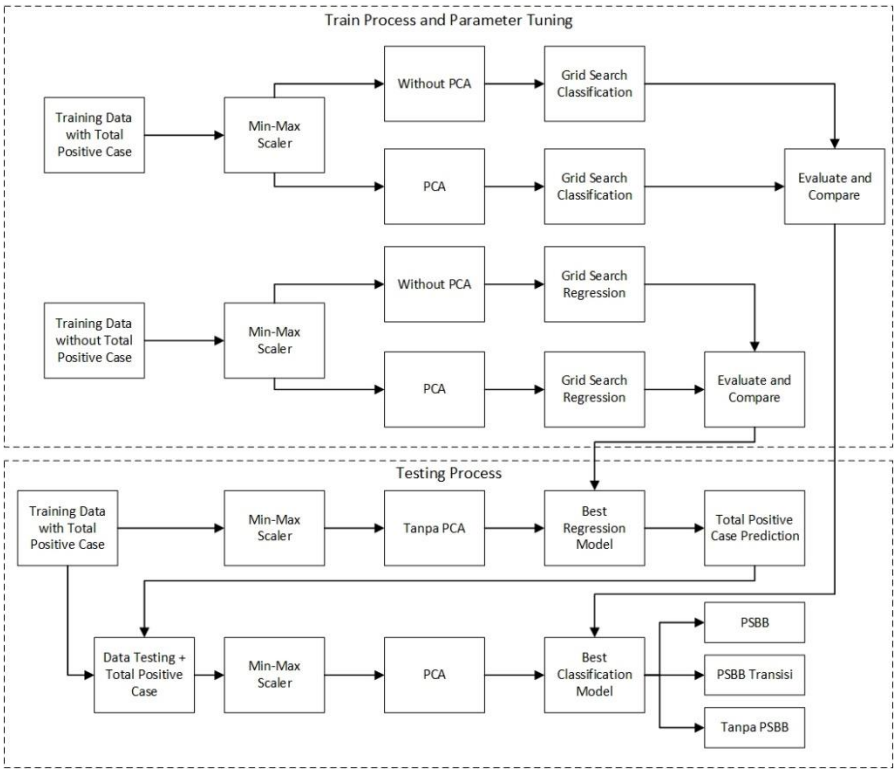


Fig. 24. Covid-19 Case Prediction and Movement Restriction Classification System Framework.

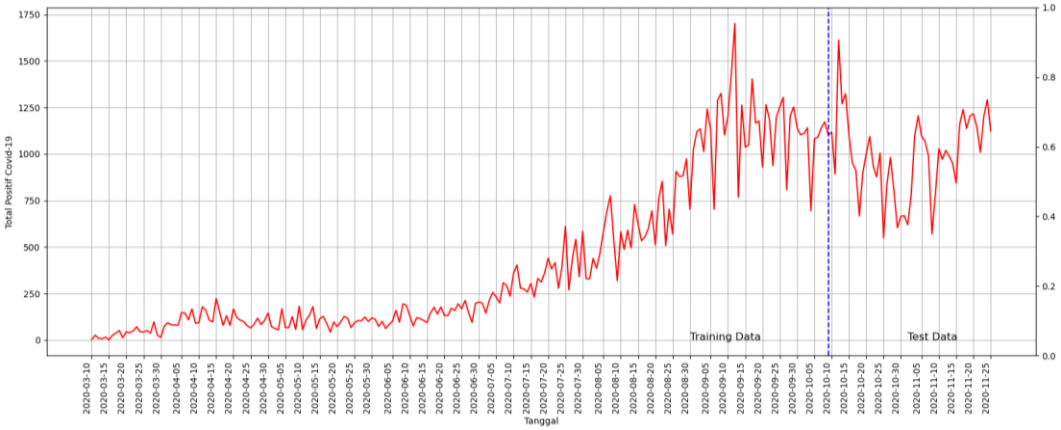


Fig. 25. Data Training and Testing Separation.

IV. Result and Discussion

We first made a regression model by comparing the evaluation results using PCA and without PCA. The Grid Search Method is used to determine the best method and the best parameters. To produce a model that can predict future values, it is necessary first to divide the training and testing data as shown in Fig. 25. The best methods, parameters, and evaluation results are shown in Table IV. Based on Table IV the best method is Multiple Linear Regression (MLR) with MAE = 148.3892 and MSE = 37036.37. If you look at the R2 value, only MLR without PCA gets a positive value. So, the MLR method without PCA was chosen as the regression method used.

Next, we need to determine a suitable method for classification, using PCA and without PCA, and determine the best parameters with grid search. The results of the grid search are shown in Table V. Based on Table V, the classification method used is a support vector machine (SVM) with PCA with a training accuracy of 0.72 and a testing accuracy of 0.8481. Because the testing accuracy exceeds 0.7, the SVM method is good enough to predict what kind of tightening will be done to the people of Jakarta. In Fig. 26, after all, forms are filled in, the system can predict the confirmed number of Covid-19 on that day totaling 871 people. In addition, we are recommended to implement the Transition PSBB by the system.

Fig. 26. System Deployment https://prediksi-covid-psbb.herokuapp.com.

TABLE IV. REGRESSION RESULT

Method	Parameter	Transforms	MAE	MSE	R2-score
MLR	**fit_intercept=True, normalize=False**	**Without PCA**	**148.3892**	**37036.37**	**0.250336**
SVR	kernel='poly', degree=3	Without PCA	536.7085	333113.5	-5.74265
Random Forest Regressor	n_estimators=100, max_depth=10	Without PCA	231.3127	77746.98	-0.5737
Decision Tree Regressor	splitter='best', max_depth=10	Without PCA	271.6456	122728.9	-1.48419
MLR	fit_intercept=True, normalize=False	PCA	307.6172	118276.6	-1.39407
SVR	kernel='poly', degree=3	PCA	866.0381	796039.1	-15.1129
Random Forest Regressor	n_estimators=50, max_depth=5	PCA	348.5216	178841.1	-2.61998
Decision Tree Regressor	splitter='best', max_depth=5	PCA	357.3924	181732.3	-2.6785

TABLE V. CLASSIFICATION RESULT

Method	Parameter	Transforms	Acc. Training	Acc. Testing
Logistic Regression	C: 10, penalty: 'l1', solver: 'liblinear'	Without PCA	0.703097	0.797468
SVM	C: 10, degree: 2, gamma: 1, kernel: 'rbf'	Without PCA	0.730237	0.772152
Decision Tree	criterion: 'gini', max_depth: 4, min_samples_leaf: 3	Without PCA	0.702732	0.670886
Naïve Bayes	alpha: 0.001, binarize: 0, fit_prior: False	Without PCA	0.489162	0.443038
KNN	algorithm: 'auto', n_neighbors: 4, p: 1, weights': 'distance'	Without PCA	0.758015	0.810127
Adaboost	algorithm: 'SAMME', learning_rate: 0.1, n_estimators: 200	Without PCA	0.713934	0.772152
XGboost	booster: 'gbtree', eta': 0.1	Without PCA	0.774499	0.810127
Logistic Regression	C: 10, penalty: 'l1', solver: 'liblinear'	PCA	0.642532	0.683544
SVM	**C: 10, degree: 2, gamma: 1, kernel: 'rbf'**	**PCA**	**0.724681**	**0.848101**
Decision Tree	criterion: 'gini', max_depth: 4, min_samples_leaf: 3	PCA	0.708379	0.683544
Naïve Bayes	alpha: 0.001, binarize: 0, fit_prior: False	PCA	0.483698	0.506329
KNN	algorithm: 'auto', n_neighbors: 4, p: 1, weights': 'distance'	PCA	0.724863	0.78481
Adaboost	algorithm: 'SAMME', learning_rate: 0.1, n_estimators: 200	PCA	0.680783	0.708861
XGboost	booster: 'gbtree', eta': 0.1	PCA	0.729964	0.734177

V. Conclusion and Future Work

In this study, the data collection and data preprocessing were explained in detail to show the reader how to collect and prepare the data before it is fed to the classification or regression methods. By detailed explanation of both the tasks, the reader can follow and develop their own data acquisition and preprocessing tasks. From the correlation analysis, it was found that the number who carry out the Covid-19 test affects the number of confirmed Covid-19 significantly. Adding data transformations such as PCA can enhance the accuracy of SVM. However, the MLR did not gain improvement. In addition, based on the test results, we were able to obtain an accuracy of 84%, which is a pretty good result. With this level of accuracy, we can be confident in using the model in actual cases. From the deployment results, we can see that the machine learning model can make predictions as we desire. The model can predict the number of Covid-19 sufferers and provide recommendations for restricting the movement of people.

Even though the accuracy of classification is relatively high, the system is still limited to the Jakarta region. The research expansion to the whole nation (Indonesia) can be more conclusive and comprehensive. The challenge will be the data collection that most of the regions do not have interactive web base covid data. So, collaboration with the government must be established. The provided data transformation is still limited to PCA. The extensive study in data transformation can be exciting as a lot of feature extraction and selection methods are available. Moreover, the technique is still limited to conventional machine learning approaches. The exploration of deep learning algorithms such as long short-term memory (LSTM) and convolutional neural networks (CNN) can be interesting discussions.

Acknowledgment

This research supported by Universitas Amikom Yogyakarta under independent research grand.

References

[1] H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," J. Autoimmun., vol. 109, no. February, pp. 18–21, 2020, doi: 10.1016/j.jaut.2020.102433.

[2] Z. Wu and J. M. McGoogan, "Characteristics of and Important Lessons from the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases from the Chinese Center for Disease Control and Prevention," JAMA - J. Am. Med. Assoc., vol. 323, no. 13, pp. 1239–1242, 2020, doi: 10.1001/jama.2020.2648.

[3] Z. E. Rasjid, R. Setiawan, and A. Effendi, "A Comparison: Prediction of Death and Infected COVID-19 Cases in Indonesia Using Time Series Smoothing and LSTM Neural Network," Procedia Comput. Sci., vol. 179, no. 2020, pp. 982–988, 2021, doi: 10.1016/j.procs.2021.01.102.

[4] N. T. P. Pang, K. Assis, M. A. Mohd Kassim, and C. M. Ho, "Analyses of the effectiveness of movement control order (Mco) in reducing the covid-19 confirmed cases in Malaysia," J. Heal. Transl. Med., vol. 24, no. Special Issue Covid-19, pp. 16–27, 2021.

[5] R. S. Pontoh et al., "Jakarta Pandemic to Endemic Transition: Forecasting COVID-19 Using NNAR and LSTM," Appl. Sci., vol. 12, no. 12, p. 5771, 2022, doi: 10.3390/app12125771.

[6] R. O. Nanda et al., "Community Mobility and COVID-19 Dynamics in Jakarta, Indonesia," Int. J. Environ. Res. Public Health, vol. 19, no. 11, p. 6671, 2022, doi: 10.3390/ijerph19116671.

[7] S. Liu et al., "Predicting the outbreak of hand, foot, and mouth disease in Nanjing, China: a time-series model based on weather variability," Int. J. Biometeorol., vol. 62, no. 4, pp. 565–574, 2018, doi: 10.1007/s00484-017-1465-3.

[8] B. Modu, N. Polovina, Y. Lan, S. Konur, A. Taufiq Asyhari, and Y. Peng, "Towards a predictive analytics-based intelligent malaria outbreakwarning system," Appl. Sci., vol. 7, no. 8, pp. 1–20, 2017, doi: 10.3390/app7080836.

[9] R. Jain, S. Sontisirikit, S. Iamsirithaworn, and H. Prendinger, "Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data," BMC Infect. Dis., vol. 19, no. 1, pp. 1–16, 2019, doi: 10.1186/s12879-019-3874-x.

[10] E. O. Nsoesie, J. S. Brownstein, N. Ramakrishnan, and M. V. Marathe, "A systematic review of studies on forecasting the dynamics of influenza outbreaks," Influenza Other Respi. Viruses, vol. 8, no. 3, pp. 309–316, 2014, doi: 10.1111/irv.12226.

[11] S. Lahmiri and S. Bekiros, "Intelligent forecasting with machine learning trading systems in chaotic intraday Bitcoin market," Chaos, Solitons and Fractals, vol. 133, 2020, doi: 10.1016/j.chaos.2020.109641.

[12] Y. Hu, C. Hu, S. Fu, P. Shi, and B. Ning, "Predicting the popularity of viral topics based on time series forecasting," Neurocomputing, vol. 210, pp. 55–65, 2016, doi: 10.1016/j.neucom.2015.10.143.

[13] O. Torrealba-Rodriguez, R. A. Conde-Gutiérrez, and A. L. Hernández-Javier, “Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models,” Chaos, Solitons and Fractals, vol. 138, 2020, doi: 10.1016/j.chaos.2020.109946.

[14] R. M. Arias Velásquez and J. V. Mejía Lara, "Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression," Chaos, Solitons and Fractals, vol. 136, 2020, doi: 10.1016/j.chaos.2020.109924.

[15] E. Buckingham-Jeffery, V. Isham, and T. House, "Gaussian process approximations for fast inference from infectious disease data," Math. Biosci., vol. 301, no. February, pp. 111–120, 2018, doi: 10.1016/j.mbs.2018.02.003.

[16] M. Yousaf, S. Zahir, M. Riaz, S. M. Hussain, and K. Shah, "Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan," Chaos, Solitons and Fractals, vol. 138, 2020, doi: 10.1016/j.chaos.2020.109926.

[17] A. N. A. Kamarudin, Z. Zainol, N. F. A. Kassim, and R. Sharif, "Prediction of COVID-19 cases in Malaysia by using machine learning: A preliminary testing," 2021 Int. Conf. Women Data Sci. Taif Univ. WiDSTaif 2021, 2021, doi: 10.1109/WIDSTAIF52235.2021.9430222.

[18] L. Di Domenico, G. Pullano, C. E. Sabbatini, P. Y. Boëlle, and V. Colizza, "Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies," BMC Med., vol. 18, no. 1, pp. 1–13, 2020, doi: 10.1186/s12916-020-01698-4.

[19] S. Panneer et al., "The Great Lockdown in the Wake of COVID-19 and Its Implications: Lessons for Low and Middle-Income Countries," Int. J. Environ. Res. Public Health, vol. 19, no. 1, 2022, doi: 10.3390/ijerph19010610.

[20] T. Sardar, S. S. Nadim, S. Rana, and J. Chattopadhyay, "Assessment of lockdown effect in some states and overall India: A predictive mathematical study on COVID-19 outbreak," Chaos, Solitons and Fractals, vol. 139, p. 110078, 2020, doi: 10.1016/j.chaos.2020.110078.

[21] N. G. Davies et al., "Association of tiered restrictions and a second lockdown with COVID-19 deaths and hospital admissions in England: a modelling study," Lancet Infect. Dis., vol. 21, no. 4, pp. 482–492, 2021, doi: 10.1016/S1473-3099(20)30984-1.

[22] B. Paital, K. Das, and S. K. Parida, "Inter nation social lockdown versus medical care against COVID-19, a mild environmental insight with special reference to India," Sci. Total Environ., vol. 728, p. 138914, 2020, doi: 10.1016/j.scitotenv.2020.138914.

[23] M. A. Acuña-Zegarra, M. Santana-Cibrian, and J. X. Velasco-Hernandez, "Modeling behavioral change and COVID-19 containment in Mexico: A trade-off between lockdown and compliance," Math. Biosci., vol. 325, no. May, p. 108370, 2020, doi: 10.1016/j.mbs.2020.108370.

[24] A. Farezi, S. A. Akbar, A. Yudhana, K. H. Ghazali, P. A. Rosyady, and A. Aminuddin, "Glucose content analysis using image processing and machine learning techniques," 2022 5th Int. Conf. Inf. Commun. Technol. ICOIACT 2022, 2022.

[25] S. A. Akbar et al., "Classification of gram-positive and gram-negative bacterial images based on machine learning algorithm," 2022 5th Int. Conf. Inf. Commun. Technol. ICOIACT 2022, 2022.

[26] E. Y. Sari, A. D. Wierfi, and A. Setyanto, "Sentiment Analysis of Customer Satisfaction on Transportation Network Company Using Naive Bayes Classifier," 2019 Int. Conf. Comput. Eng. Network, Intell. Multimedia, CENIM 2019 - Proceeding, vol. 2019-Novem, Nov. 2019, doi: 10.1109/CENIM48368.2019.8973262.

[27] A. Aminuddin and F. Ernawan, "AuSR1: Authentication and self-recovery using a new image inpainting technique with LSB shifting in fragile image watermarking," J. King Saud Univ. - Comput. Inf. Sci., Feb. 2022, doi: 10.1016/J.JKSUCI.2022.02.009.

[28] A. Aminuddin and F. Ernawan, "AuSR2: Image watermarking technique for authentication and self-recovery with image texture preservation," Comput. Electr. Eng., vol. 102, p. 108207, Sep. 2022, doi: 10.1016/J.COMPELECENG.2022.108207.

[29] F. Ernawan, A. Aminuddin, D. Nincarean, M. F. A. Razak, and A. Firdaus, "Three Layer Authentications with a Spiral Block Mapping to Prove Authenticity in Medical Images," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 4, 2022, doi: 10.14569/IJACSA.2022.0130425.

Evaluation of Parameter Fine-Tuning with Transfer Learning for Osteoporosis Classification in Knee Radiograph

Usman Bello Abubakar[1]
Computer Science
Baze University
Abuja, Nigeria

Moussa Mahamat Boukar[2]
Computer Science
Nile University of Nigeria
Abuja, Nigeria

Steve Adeshina[3]
Computer Engineering
Nile University of Nigeria
Abuja, Nigeria

***Abstract*—Osteoporosis is a bone disease that raises the risk of fracture due to the density of the bone mineral being low and the decline of the structure of bone tissue. Among other techniques, such as Dual-Energy X-ray Absorptiometry (DXA), 2D x-ray pictures of the bone can be used to detect osteoporosis. This study aims to evaluate deep convolutional neural networks (CNNs), applied with transfer learning techniques, to categorize specific osteoporosis features in knee radiographs. For objective labeling, we obtained a selection of patient knee x-ray images. The study makes use of the Visual Geometry Group Deep (VGG-16), and VGG-16 with fine-tuning. In this work, the deployed CNNs were assessed using state-of-the-art metrics such as accuracy, sensitivity, and specificity. The evaluation shows that fine-tuning enhanced the VGG-16 CNN's effectiveness for detecting osteoporosis in radiographs of the knee. The accuracy of the VGG-16 with parameter fine-tuning was 88% overall, while the accuracy of the VGG-16 without parameter fine-tuning was 80%.**

Keywords*—Osteoporosis; transfer learning models; convolutional neural network; fine-tuning

I. Introduction

Osteoporosis is a severe illness common in about 9% of citizens, above 50 years, in the United States [1] and about 200 million women worldwide. One in three people in developed nations may experience an osteoporotic compression fracture (OCF) [1]. The likelihood of recurrent fractures greatly increases after the initial fracture [2] [3] [4]. Even one OCF is linked to a greater death rate and a lower quality of life [5].

Osteoporosis, which is defined as porous bone, is a condition in which the mass of the bone is low and the bone tissues have undergone microarchitectural deterioration. Osteoporosis increases fracture risk of the wrist, hip, and spine, among other bones, and lowers bone mineral density (BMD). Additionally, osteoporosis alters the quantity and type of proteins in bones. Osteoporotic fractures are described as those that happen at a site where there is low BMD and are more likely to happen beyond the age of roughly 50 [6] [7].

Every individual irrespective of gender and race could be affected by the disease and as the population ages, its prevalence would also increase. Among specialists, it is known as a silent bone disease because its symptoms are not spotted before a fracture and thus, pose threats to a patient by inducing other secondary bone problems like arthritis and the likes [8]. In the skeletal system, there is a continuous activity of bone tissues been lost by resorption, and also bone tissues have been rebuilt back by formation. The system is said to be at a bone loss when bone tissue formation is less than bone tissue resorption [9].

It has long been believed that deep learning is effective at learning feature categorization from medical images [10]. Deep Learning (DL) classifiers utilize high-dimensional features to improve the performance of DL networks in object detection and image classification. Machine Learning (ML) techniques, in contrast to DL techniques, rely on explicitly categorized features [11].

Deep CNNs have been proved to be efficient tools for categorizing images, but they are difficult to employ with medical radiographic image data since they require a large amount of training data. Transfer learning is recognized as an efficient method in training deep CNNs when the dataset is small to prevent overfitting [12].

We use a dataset of knee radiographs (or knee X-rays) to apply and assess deep transfer learning algorithms for classifying osteoporosis. This work objectively assessed the impact of parameter fine-tuning on a transfer learning deep CNN model's performance for identifying knee radiograph pictures based on the BMD value (T-score).

II. Related Work

Authors in [13] performed a comparison of classification systems for osteoporosis prediction using feature selection based on wrappers. As classification methods, multilayer feed-forward neural network (MFNN), Naive Bayes, and logistic regression were employed. Single Nucleotide Polymorphisms (SNPs), age, menopause, and BMI of Taiwanese women were all included in the dataset utilized for the study.

The three classifiers, utilizing SNP, were tested using a 10-fold cross-validation method both with feature selection and without feature selection. Without using wrapper-based feature selection, the Area under Curve (AUC) for the MFNN was 0.489. The AUC for naive Bayes was 0.462, and the AUC for logistic regression prediction was 0.485 [13].

The performance metric for classifiers utilizing a wrapper-based strategy yielded an AUC of 0.631 for MFNN, AUC of 0.569 naïve Bayes, and AUC of 0.620 for logistic regression models [13]. The experimental results demonstrated that the MFNN model with the wrapper-based technique was the most accurate predictive model for predicting disease susceptibility in Taiwanese women based on the complicated interplay between osteoporosis and SNPs. The findings reveal that the proposed technology can help patients and clinicians make better decisions based on clinical data such as SNP genotyping data [13].

The study proposed by [14] investigates whether adding clinical information improves diagnosis when compared to images alone when using deep learning. 1131 images from patients who had skeletal bone mineral density testing and hip radiography at the same general hospital between 2014 and 2019 were gotten. From hip radiographs, five convolutional neural networks (CNN) models were employed to assess osteoporosis [14]. Adding clinical values increased accuracy, sensitivity, and specificity.

Using only hip radiograph images, without clinical covariates, GoogleNet and EfficientNet b3 models displayed the highest levels of model performance. EfficientNet b3 demonstrated the best accuracy, sensitivity, and other metric core among the five ensemble models when patient factors were taken into account [14]. Increasing clinical covariates increased the accuracy of the deep learning models [14].

The authors in [15] revealed that dental panoramic radiographs can be used to accurately diagnose osteoporosis using CNNs. Additionally, integrating patient factors in common clinical contexts enhanced all predictions' performance measures in comparison to using the image-only mode. The study hypothesized that advanced inference, which is possible by deep learning, which, in turn, simultaneously takes important information about clinical factors into account that cannot be determined from dental panoramic X-ray images alone, led to an increase in diagnosis precision [15].

Various implementations of EfficientNet and ResNet were employed in the study by the authors. The most accurate ResNet and EfficientNet techniques, respectively, were ResNet-152 and EfficientNet-B7. However, EfficientNet-b7 obtained better results than other CNN models [15]. Gradient-weighted Class Activation Mapping (Grad-CAM) was used to visualize learning. ResNet concentrated on the cortical bone at the base of the jaw. Contrarily, EfficientNet concentrated on the area above the cortical bone as well as the cortical bone at the bottom border of the jaw.

The authors in [16] developed a cutting-edge, reliable bone disease prediction model based on recognized risk factors. Then it was feasible to discover the early risk factors for determining the beginnings of bone disorders using Pre-training and fine-tuning. The most significant risk factors are coupled with model parameters during the pre-training phase to calculate contrastive divergence, which minimizes record size.

Using the ground truth values "g1" and "g2," where "g1" stood for osteoporosis and "g2" for a rate of bone loss, the outcomes of the preceding phase were compared [16]. The model was produced using a Deep Belief Network (DBN), and it was then contrasted with models made both before and after essential feature identification. The study's conclusions indicated that adding pertinent variables might improve the predictive model's performance.

The authors in [17] built a model to predict the risk of osteoporosis using supervised machine learning. The study made public the variables that experts considered while determining the risk of osteoporosis. Developing a predictive model for the identification of people in Nigeria who are at risk for osteoporosis was the study's main objective. The supervised machine learning techniques Nave Bayes (NB) classifier and Multi-layer Perceptron were utilized to develop the predictive model for osteoporosis risk (MLP). The identification and data collection from patients in Nigerian hospitals found that there were 20 risk markers, including CD4 count levels classified as low, moderate, and high risk [17]. According to their finding, NB got 71.4% accuracy while the MLP had the best got 100%.

There has been a scarcity in the use of DL to interpret and predict osteoporosis from a knee radiograph. This research aims at filling this gap in the existing knowledge that points to the need for further understanding and investigation of osteoporosis prediction using DL from knee radiographic images.

III. METHODS

A. Research Design

This research tries to classify osteoporosis in knee radiographs. To replicate the osteoporosis diagnostic range in the DXA approach, we employed a segmented dataset. In addition, the Keras Deep Learning (DL) packages were employed for data normalization and augmentation. The diagnosis of osteoporosis from knee radiographs was performed using the VGG-16 transfer learning deep neural network. We examined the accuracy of the osteoporosis prognostic diagnostic using the transfer learning model with and without parameter fine-tuning using cutting-edge performance metrics.

B. Dataset

The dataset, published in August 2021, was gotten from Mendeley data uploaded by [18]. The dataset images were statistically augmented (i.e. increased) using data augmentation in python. Fig. 1 shows two images from the dataset indicating osteoporosis cases and normal cases.

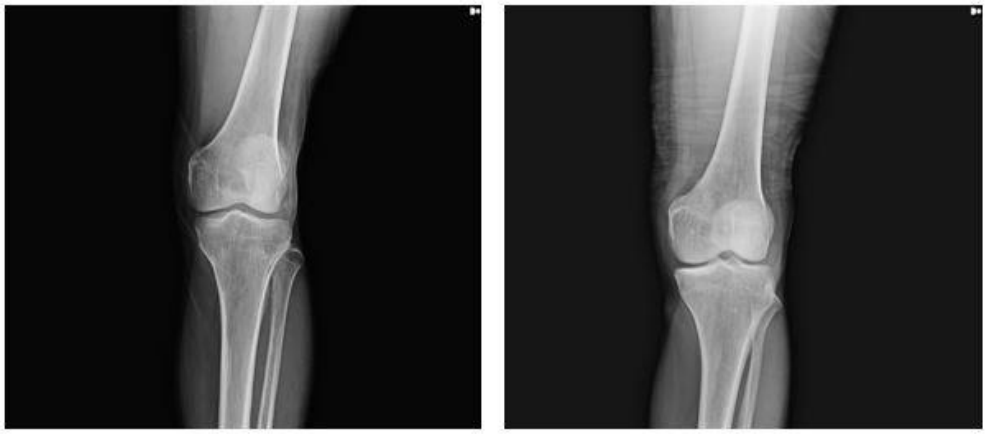


Fig. 1. Osteoporosis Case and Normal Case [18].

The dataset, after statical augmentation using python augmentation functions, comprises 323 normal knee radiograph images and 323 osteoporotic knee radiograph images of patients. Table I shows the splitting of image data into train, test, and validation data.

TABLE I. IMAGE DISTRIBUTION

Class	Total	Training	Testing
Normal (0)	323	259	65
Osteoporosis (1)	323	259	65

C. Grayscale Conversion

The dataset consists of images in Red Green Blue (RGB) format. A three-dimensional byte array (i.e., RGB image) stores a color value for each pixel. RGB format increases the complexity of training the model. Grayscale (i.e., black and white images) are preferred as they simplify computational complexity.

The modality of our research is based on knee x-ray data and thus, in an x-ray, color is irrelevant to diagnosis. Due to this reason, and the fact that grayscale images are easier to train a deep learning network, the images were converted from RGB to grayscale using the OpenCV python library.

D. Data Normalization

It is the process of converting image data pixels to a predetermined range : (0, 1) or (-1, 1). The pixel values in most images range from 0 to 255. Training a deep neural network with large integer values can interfere with or slow down the learning process. Therefore, picture normalization is a recommended practice: pixel values range between 0 and 1.

The images in the dataset were normalized (rescaled) using the python ImageDataGenerator method and passing rescale=1. /255 as its argument.

E. Data Augmentation

When working with deep learning models, it is paramount to ensure that the model gets a sufficient amount of training data. Data augmentation is the application of various changes to original images, resulting in several altered copies of the same image. Each replica, however, differs from the others in some ways due to the augmentation procedures used.

For this study, augmentation was done using Keras ImageDataGenerator in python. Itemized below are some of the techniques applied:

1) Standardization
2) Rotation
3) Shifts
4) Brightness changes, among others

The Keras ImageDataGenerator class is intended to give real-time data augmentation, which is said to be its key advantage. Every epoch, the model is given fresh versions of the images due to the ImageDataGenerator python class.

F. Transfer Learning Model Used (VGG-16)

In this work, two CNN study groups were used: VGG16 and the parameter fine-tuning model from VGG16. The difference between the two implementations is that the latter used parameter fine-tuning while the former did not. This was performed by unfreezing a couple of the original model's top levels and training the newly added classifier layers alongside the base model's final layers. The schematic diagram for the two transfer learning models used in this work is depicted in the block diagrams in Fig. 2 and Fig. 3.

The VGG16 architecture was chosen since it had been widely adopted and considered cutting-edge in image classification applications trained on a large dataset [10] [19].

G. Training the Model

Five folds were randomly selected from the training dataset of the chosen images. This prevented bias or overfitting while performing a five-fold cross-validation on the model training. The dataset was split into independent training and validation sets within each fold using an 80 to 20 split. A validation set that was completely different from the other training folds was chosen to assess the training state throughout training. Once one model training phase was complete, the other independent fold was utilized as a validation set, and the previous validation set was recycled as part of the training set to evaluate the model training. Fig. 4 shows a five-fold cross-validation done in this study.

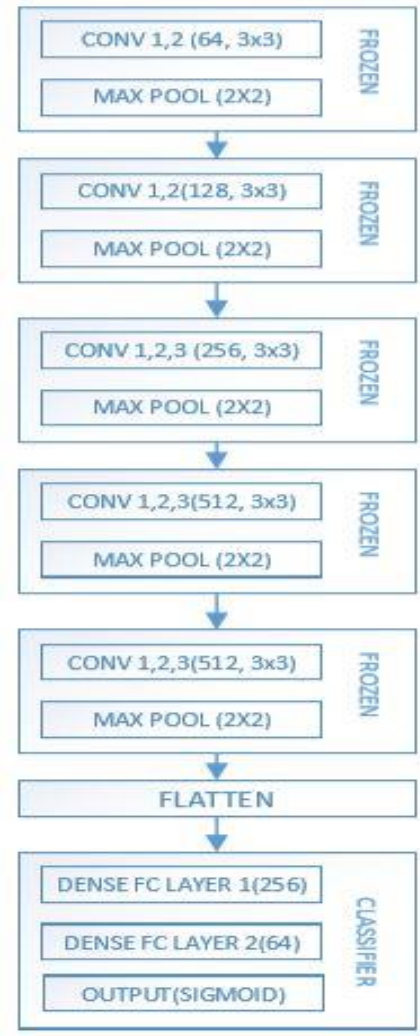


Fig. 2. VGG-16 without Parameter Fine-Tuning.

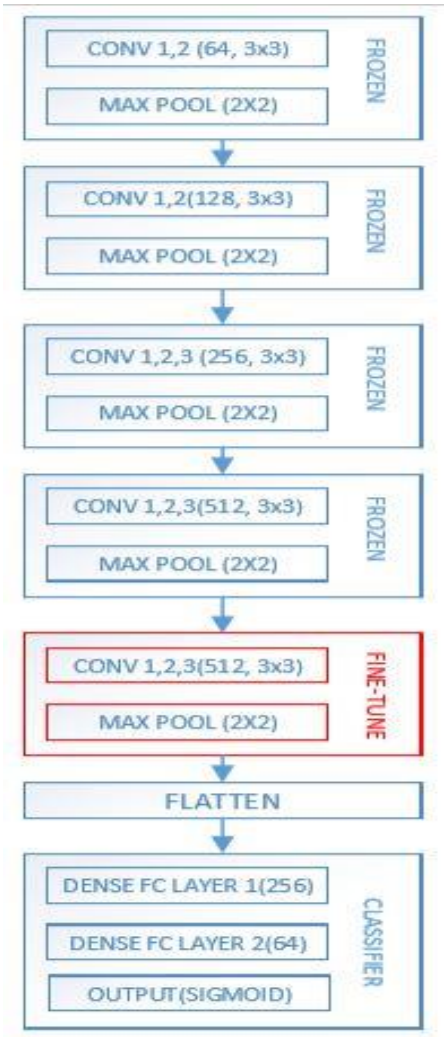


Fig. 3. VGG-16 with Parameter Fine-Tuning.

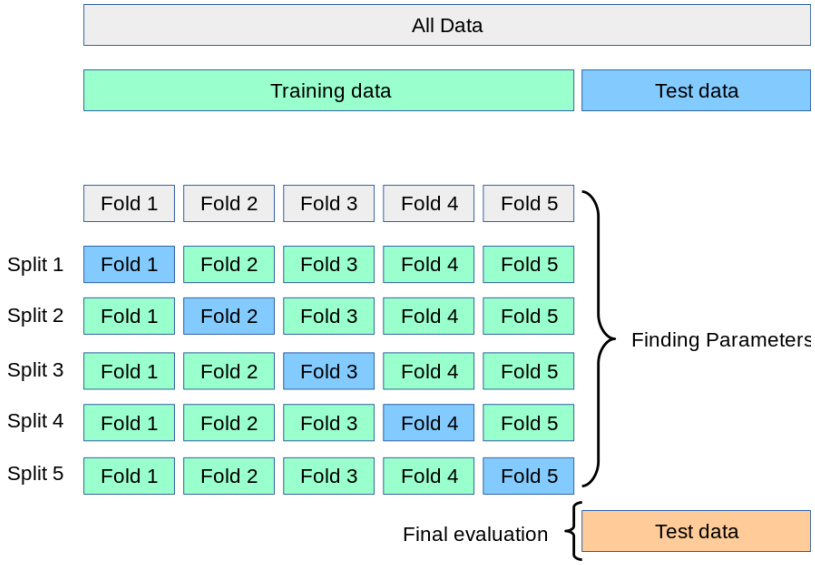


Fig. 4. Overview of 5-Fold Cross Validation.

This process of cross-validation was repeated for the VGG-16 without parameter fine-tuning and for the VGG-16 with parameter fine-tuning. The Google colabs Graphics Processing Unit (GPU) was used to train and test all models. The Keras library and TensorFlow were used throughout the process of applying the transfer learning deep learning models.

IV. RESULTS

A. Performance Metrics

The following metrics were established for each model to fully assess its performance: (1) sensitivity, (2) specificity, (3) accuracy, (4) precision, and (5) F1-score. The formula for the specified metrics is expressed below.

$$\text{Sensitivity} = \frac{true\ positive}{true\ positive + false\ negative} \quad (1)$$

$$\text{Specificity} = \frac{true\ negative}{true\ negative + false\ negative} \quad (2)$$

$$\text{Accuracy} = \frac{true\ negative + true\ positive}{all\ cases} \quad (3)$$

$$\text{Precision} = \frac{true\ positive}{true\ positive + false\ positive} \quad (4)$$

$$\text{F1} = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

B. Confusion Matrix

A method for summarizing a classification algorithm's performance is the confusion matrix (CM). In addition to giving insight into the mistakes the classifier is making, it also reveals the specific mistakes that are occurring. The confusion matrix helps to overcome the limitation of using classification accuracy alone. Fig. 5 and Fig. 6 show the confusion matrix for the VGG-16 model without parameter fine-tuning and the VGG-16 model with parameter fine-tuning respectively.

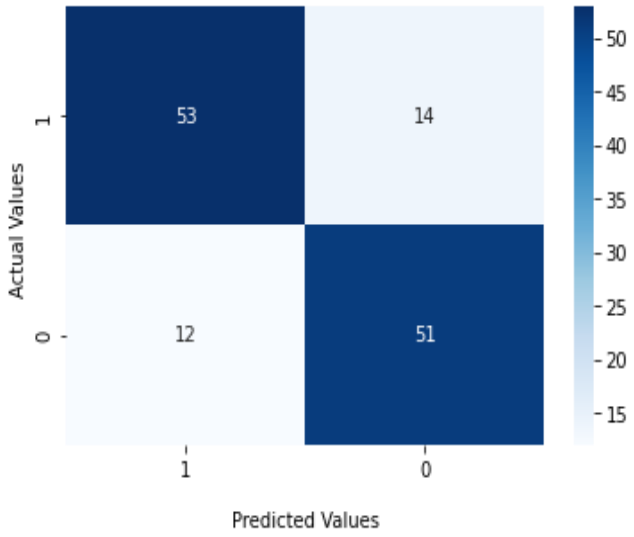


Fig. 5. Confusion Matrix for VGG-16 without Parameter Fine-Tuning.

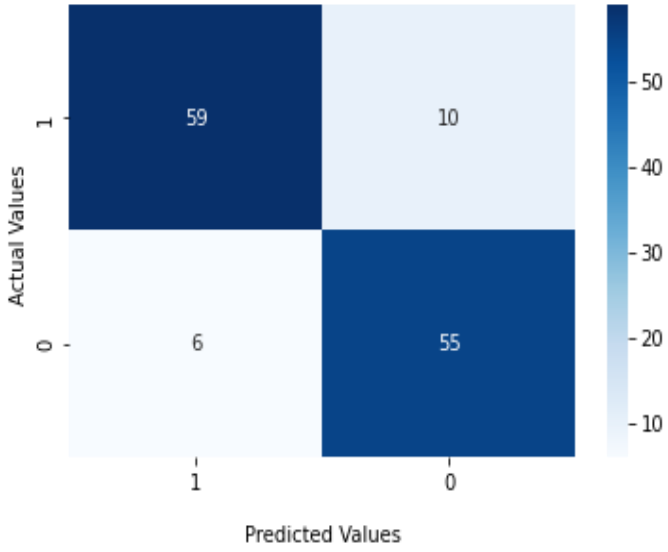


Fig. 6. Confusion Matrix for VGG-16 with Parameter Fine-Tuning.

C. Prediction Performance

The osteoporosis patient knee x-rays dataset has been tested using the CNN models utilized in this work. The dataset was divided into training and testing portions in an 80:20 ratio for all transfer learning models. The overall accuracy obtained for the two classifiers on the dataset is summarized in Table II. Each model underwent 50 epochs of training. For all models, as the loss metric, binary_crossentropy was used as the dataset target has two classes (i.e., binary classification problem). RMSprop is the chosen optimizer, and its learning rate is 0. 001.

The Keras evaluate function was invoked on the compiled model with the test data as an argument to evaluate the accuracy of the models. Table III provides a comparison of our work with similar works. Fig. 7 shows a chart visually depicting the performance difference between the two implementations of the VGG-16 transfer learning model.

D. Algorithm Justification

The justification for choosing VGG-16 architecture was that it had been widely adopted and recognized as state-of-the-art in both general and medical image classification tasks but has not readily been applied to osteoporosis classification from patient knee radiographs. Additionally, VGG-16 has been trained on large-scale datasets, so that a transfer learning approach could be adopted for large-scale image recognition.

The reason for using parameter fine-tuning is that research shows it boosts the performance of a deep learning model over random initialization [20].

E. Dataset Justification

The reason for choosing the knee radiograph dataset is because deep learning research on osteoporosis classification using knee x-ray is still relatively scarce.

F. Limitations of the Study

A deep learning model requires massive amounts of data to be efficient. The number of training observations in the dataset was not large enough and hence poses a limitation to the study. However, data augmentation was applied to mitigate such limitations.

G. Recommendation and Future Work

The perception based on the findings stipulates that overfitting in transfer learning due to few data samples can be avoided using certain techniques: cross-validation, data augmentation, and parameter fine-tuning. Findings also show that parameter fine-tuning in transfer learning can be used to significantly increase the accuracy, sensitivity, specificity, precision, and F1 of a deep learning model.

Osteoporosis is caused not just by low bone mineral density, but also by other factors such as age, gender, weight, height, and so on. These are clinically important risk factors for osteoporosis. For future work, we would like to extend our methods by adding patient variables such as age, and gender, amongst others, as clinical covariates to create an ensemble model with the transfer learning models

TABLE II. RESULTS OBTAINED

	Ac	Se/Re	Sp	Pr	F1
VGG-16 without Fine-Tuning	0.80	0.82	0.81	0.79	0.80
VGG-16 with Fine-Tuning	0.88	0.91	0.90	0.86	0.88

*AC: ACCURACY, SE: SENSITIVITY, RE: RECALL, SP: SPECIFICITY, PR: PRECISION

TABLE III. COMPARISON WITH OTHER WORKS

	Classifier	Accuracy	Sensitivity/Recall	Specificity
Our Paper	VGG-16	0.80	0.82	0.81
Our Paper	VGG-16: Fine-Tuning	0.88	0.91	0.90
[14]	ResNet-18	0.79	0.86	0.86
[12]	CNN with 3 layers	0.66	0.68	0.65
[15]	ResNet-50	0.83	0.75	0.90

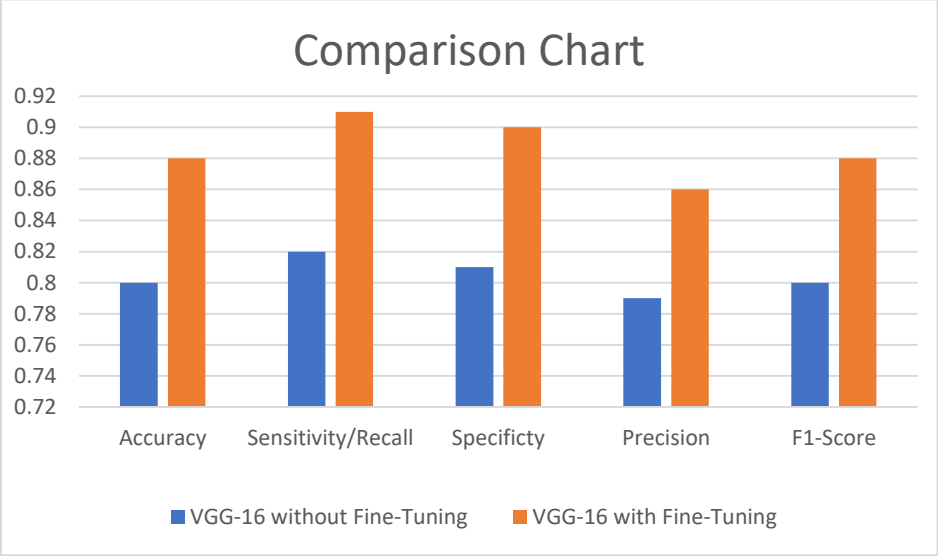


Fig. 7. Comparison Chart.

V. CONCLUSION

In circumstances when there is a small training dataset, this study demonstrates the efficacy of deep CNN tuning and transfer learning for detecting osteoporosis in knee x-ray images. On networks that have already been trained for the categorization of osteoporosis, we have used the VGG-16 transfer learning technique. According to the experimental findings, the fine-tuning technique enabled transfer learning to obtain an overall accuracy of 88%, which was higher than that of 80% achieved by transfer learning without fine-tuning.

The results show that parameter fine-tuning in transfer learning can be used to significantly increase the accuracy, sensitivity, specificity, precision, and F1 of a deep learning model. For future work, we would like to extend our methods by creating an ensemble approach of adding patient clinical covariates to classify osteoporosis with VGG-16 from knee radiograph.

This research was broken into several parts: Introduction, related works, methods, results, and conclusion. The method section provided details as to how the dataset was acquired, the augmentation techniques used, the grayscale conversion of images from RGB to grayscale, the cross-validation split used, and the transfer learning model applied. The results section depicted some state-of-the-art deep learning evaluation metrics used to evaluate the transfer learning variations of the VGG-16 model used.

REFERENCES

[1] N. C. Wright et al., "The recent prevalence of osteoporosis and low bone mass in the United States based on bone mineral density at the femoral neck or lumbar spine," Journal of Bone and Mineral Research, vol. 29, no. 11, pp. 2520–2526, Nov. 2014, doi: 10.1002/jbmr.2269.

[2] A. B. Hodsman, W. D. Leslie, J. F. Tsang, and G. D. Gamble, "10-year probability of recurrent fractures following wrist and other osteoporotic fractures in a large clinical cohort: an analysis from the Manitoba Bone Density Program," Arch Intern Med, vol. 168, no. 20, pp. 2261–2267, Nov. 2008, doi: 10.1001/ARCHINTE.168.20.2261.

[3] S. Roux et al., "The World Health Organization Fracture Risk Assessment Tool (FRAX) underestimates incident and recurrent fractures in consecutive patients with fragility fractures," J Clin Endocrinol Metab, vol. 99, no. 7, pp. 2400–2408, 2014, doi: 10.1210/JC.2013-4507.

[4] C. M. Robinson, M. Royds, A. Abraham, M. M. McQueen, C. M. Court-Brown, and J. Christie, "Refractures in patients at least forty-five years old. a prospective analysis of twenty-two thousand and sixty patients," J Bone Joint Surg Am, vol. 84, no. 9, pp. 1528–1533, 2002, doi: 10.2106/00004623-200209000-00004.

[5] J. R. Center, T. v. Nguyen, D. Schneider, P. N. Sambrook, and J. A. Eisman, "Mortality after all major types of osteoporotic fracture in men and women: an observational study," Lancet, vol. 353, no. 9156, pp. 878–882, Mar. 1999, doi: 10.1016/S0140-6736(98)09075-8.

[6] J. A. Kanis, A. Oden, O. Johnell, C. de Laet, B. Jonsson, and A. K. Oglesby, "The components of excess mortality after hip fracture," Bone, vol. 32, no. 5, pp. 468–473, 2003, doi: 10.1016/S8756-3282(03)00061-9.

[7] O. Johnell and J. A. Kanis, "An estimate of the worldwide prevalence and disability associated with osteoporotic fractures," Osteoporosis International, vol. 17, no. 12, pp. 1726–1733, Dec. 2006, doi: 10.1007/S00198-006-0172-4.

[8] T. Sozen, L. Ozisik, and N. Calik Basaran, "An overview and management of osteoporosis," European Journal of Rheumatology, vol. 4, no. 1, pp. 46–56, Mar. 2017, doi: 10.5152/EURJRHEUM.2016.048.

[9] B. L. Riggs et al., "Changes in bone mineral density of the proximal femur and spine with aging. Differences between the postmenopausal and senile osteoporosis syndromes," Journal of Clinical Investigation, vol. 70, no. 4, pp. 716–723, 1982, doi: 10.1172/JCI110667.

[10] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/J.MEDIA.2017.07.005.

[11] U. Bello Abubakar, M. Mahamat Boukar, and S. Dane, "Review of Swarm Fuzzy Classifier and a Convolutional Neural Network with VGG-16 Pre-Trained Model on Dental Panoramic Radiograph for Osteoporosis Classification", Accessed: Jul. 26, 2022. [Online]. Available: www.jrmds.in

[12] K. S. Lee, S. K. Jung, J. J. Ryu, S. W. Shin, and J. Choi, "Evaluation of Transfer Learning with Deep Convolutional Neural Networks for Screening Osteoporosis in Dental Panoramic Radiographs," Journal of Clinical Medicine, vol. 9, no. 2, Feb. 2020, doi: 10.3390/JCM9020392.

[13] H. W. Chang, Y. H. Chiu, H. Y. Kao, C. H. Yang, and W. H. Ho, "Comparison of classification algorithms with wrapper-based feature selection for predicting osteoporosis outcome based on genetic factors in a Taiwanese women population," International Journal of Endocrinology, vol. 2013, 2013, doi: 10.1155/2013/850735.

[14] N. Yamamoto et al., "Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates," Biomolecules, vol. 10, no. 11, pp. 1–13, Nov. 2020, doi: 10.3390/BIOM10111534.

[15] S. Sukegawa et al., "Identification of osteoporosis using ensemble deep learning model with panoramic radiographs and clinical covariates," Scientific Reports 2022 12:1, vol. 12, no. 1, pp. 1–10, Apr. 2022, doi: 10.1038/s41598-022-10150-x.

[16] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial Intelligence, vol. 97, no. 1–2, pp. 273–324, 1997, doi: 10.1016/S0004-3702(97)00043-X.

[17] E. N. Chidozie et al., "Osteoporosis Risk Predictive Model Using Supervised Machine Learning Algorithms," http://www.sciencepublishinggroup.com, vol. 5, no. 6, p. 78, Jan. 2018, doi: 10.11648/J.SR.20170506.11.

[18] I. Majeed Wani and S. Arora, "Knee X-ray Osteoporosis Database," vol. 2, 2021, doi: 10.17632/FXJM8FB6MW.2.

[19] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/S11263-015-0816-Y.

[20] "How transferable are features in deep neural networks? | Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2." https://dl.acm.org/doi/10.5555/2969033.2969197 (accessed Jan. 10, 2022).

Dangerous Goods Container Location Allocation Strategy based on Improved NSGA-II Algorithm

Xinmei Zhang[1]
College of Mechanical and Electronic Engineering
China University of Petroleum
Shandong Qingdao, China

Nannan Liang[2]
College of Economy and Management
China University of Petroleum
Shandong Qingdao, China

Chen Chen[3]
College of Economy and Management
China University of Petroleum
Shandong Qingdao, China

***Abstract*—The characteristics of port dangerous goods are complicated and diverse in danger, which is very likely to cause chain effects once a fire and explosion accident occurs. Based on the distribution characteristics of dangerous goods container yards and the special national storage requirements for dangerous goods containers, the paper establishes a multi-objective optimization model with a double priority of safety and economy, starting from reducing the number of reversals. The improved non-dominated sorting genetic algorithm based on the elite strategy was used to solve the model and the algorithm was tested and improved. Based on the Pareto optimal solution set, the entropy weight-TOPSIS method was used to optimize the sorting of multiple solution sets, which improved the performance of the algorithm. The analysis further clarifies the important relationship between attributes, and the running time is shortened by 85.7% compared with the traditional NSGA algorithm. The optimization model and algorithm can provide decision support for the actual operation and management of container storage, and provide a good reference for accident risk prevention and control.**

***Keywords*—Dangerous goods containers; container allocation; improved NSGA-II algorithm; entropy weight-TOPSIS**

I. Introduction

The wide range of chemicals in port, hazardous chemical yards, their high hazardousness, and mobility poses a double challenge for business process optimization and risk control in the yards. In 2015, the "8.12" fire and explosion accident in Tianjin Port caused serious casualties and property damage, which was mainly caused by the irregular management of dangerous goods container yard storage, and the serious phenomenon of overloading, over-height, and irregular mixed storage. To further strengthen the safety management of dangerous goods in ports and to prevent and reduce dangerous goods accidents, the Ministry of Transport of the People's Republic of China revised the Provisions on the Safety Management of Dangerous Goods in Ports in 2017, putting forward higher requirements on the risk management of dangerous goods in ports.

The allocation of container yard space is a key factor that restricts the efficiency of terminals and increases the operational costs of the terminal. Reducing the number of unloads to increase terminal efficiency and reduce operating costs has become the focus of research in this field. Zhang and Ambrosino et al. [1], [2] consider the weight of the container and develop a dynamic model to reduce the number of reversals. Galle et al. [3] established the optimization model of container pre-marshaling by considering the loading order. A new unified integer programming model was designed to solve the problem of reducing the number of containers unloaded. In addition to the weight and shipping order, the uncertainty of delivery time [4], exit box entry time [5], [6], and pick-up time [7], [8] are also critical factors that affect unloading operations. Besides the number of box dumps, the task allocation of the bridge [9] and moving path [10] also affect the box allocation strategy. All the strategies mentioned are used for ordinary containers, with no consideration of dangerous goods container stowage rules, Zhou et al. [11] established a distribution optimization model for dangerous goods containers by considering the storage height limit, and this model was solved using the Monte Carlo tree algorithm, it improves the efficiency of putting boxes away, but it does not take into account the impact of the number of reversals on safety.

Many algorithms to solve bin allocation exist, such as the heuristic algorithm [6], [12] particle swarm algorithm [13], tabu search algorithm [14], mixed harmony simulated annealing algorithm [10], [15], and genetic algorithm [9], [16], [17]. Among these, the genetic algorithm is used more widely used. Tang et al. [18] designed a genetic algorithm-based heuristic to solve the storage problem of a large iron ore terminal; Jun and Chen [19] established a mixed-integer programming model based on yard crane resource optimization and used a genetic algorithm to solve it. Based on the above studies, it can be found that genetic algorithms are used more than the other types and can better solve the problem of bin allocation. However, most of the previous studies used traditional genetic algorithms and traditional NSGA-II algorithms, which lack a diversity of solutions. Therefore, this research focuses on the control of the elite range in the algorithm design and proposes an improved NSGA-II algorithm to improve the diversity of solutions and realize the convergence of the search algorithm to the global optimal solution.

In summary, it can be seen that the current research on the allocation of container yard space is mostly directed at ordinary containers, with the core objective of improving operational efficiency and lacking attention to the safety of dangerous goods yards. Taking into account the special characteristics of dangerous goods containers and the efficiency requirements of

storage operations, the storage process should be regarded as a multi-objective optimization problem with the double priority of safety and economy.

This research mainly takes dangerous goods containers as the research object and establishes a storage yard optimization model based on safety and economic benefits. At the same time, in order to further improve the diversity of solutions and realize the convergence of the search algorithm to the global optimal solution, it focuses on the control of the elite range and proposes an Improved NSGA-II algorithm. Then entropy weight-TOPSIS sorting method was used to conduct the multi-attribute decision-making analysis and the Pareto optimal solution is obtained, to obtain the optimal solution to further reduce the storage risk of dangerous cargo containers and improve the operation efficiency.

II. Mathematical Model

A. Problem Description

A dangerous goods storage yard is where Dangerous goods containers are stored. Therefore, once an emergency occurs, the hazard is extremely high. The safety of hazardous chemical container yards is embodied in the following three aspects: classified storage, number of container dumps, and storage height. The required storage height depends on the type of dangerous goods. According to the "Safety Regulations for Port Operation of Hazardous Chemicals Containers" flammable and explosive Dangerous goods containers should only be stacked up to two tiers, and other Dangerous goods containers shall not exceed three tiers. Moreover, effective isolation should be prepared according to the nature of the dangerous goods. Generally, since the storage height is low, the movement of Dangerous goods containers employs manual truck operations, which do not involve the problem of the yard and bridge schedules. Most hazardous chemical container yards in ports have been zoned according to the isolation requirements of hazardous chemicals. Therefore, this study focused only on Dangerous goods containers with fixed zones.

From a safety perspective, reducing unnecessary container handling operations and minimizing the storage height can reduce the crane workload, prevent stacks from being overly high, and dumping over to reduce the safety hazards caused by Dangerous goods containers during operation. From an economic perspective, reducing the operation of unloading can increase the efficiency of yard operation and reduce costs. The exit time is known for containers entering the hazardous chemical container yard. When a container enters the yard, it is necessary to optimize the bin allocation sequence to reduce the unloading operation. The exit time was early on the top floor. Generally, heavy containers are placed on the lower layer to ensure safety when shipping containers, and lighter boxes are placed on the upper layer. Therefore, when allocating bin positions in the yard, it is common to place heavy boxes on the upper layer and lighter boxes on the lower layer.

The problem of optimizing the allocation of Dangerous goods containers can be summarized as minimizing the operation of unloading containers in a range of storage yards when the number of bays, rated height, and initial storage status is known. Moreover, the order of appearance, the weight of the box, and the height of the stack also affect the unloading operation. Therefore, it should be considered when establishing the optimization model.

B. Model Assumption

Based on the nature of the hazardous chemical container and storage yard scenario, the following assumptions were made to achieve the goal of optimizing storage:

The loading and unloading equipment are fault-free, and all the operation links are normal.

All the boxes on-site meet the isolation requirements for Dangerous goods containers.

Only for Dangerous goods containers.

The type of dangerous goods and the quality, size, and time of entry and exit of the dangerous chemical container are known.

C. Notations and Variables

I: The set of all containers, $I = \{1,2,3, \dots, N_i\}$, $i, j \in I$;

N_i: Total container arrivals;

K : The set of all fields, $K = \{1,2,3, \dots, N_k\}$, 1 represents $K1$, 2 Represents $K2$, 3 represents $K3$, 4 represents $K4$, 5 represents $K5$, 6 re $c$presents $K3$, 7 represents $K8$, 8 represents KD, 9 represents KE, 10 represents KF, 11 represents $N1$, 12 represents $N2$, 13 represents $N3$, 14 represents $N4$, 15 represents $N5$, 16 represents $N6$;

B: The set of all shells, $B = \{01,02,03, \dots, N_b\}$;

R: The set of all rows, $R = \{1,2,3, \dots, N_r\}$;

T: The set of all layers, $T = \{1,2,3, \dots, N_t\}$

Q : The set of all container locations, $Q = \{10111,10111, \dots, 100111,100112, \dots, N_k \times 1000 + N_b \times 100 + N_r \times 10 + N_t\}$;

s_i : 1 if put into the designated field according to the category, 0 otherwise;

$E_{i,j}$: 1 if the container in the first field is on the lower level, 0 otherwise;

$x_{i,b,r,t}$: 1 if container i is placed in the container spaces (b, r, t), 0 otherwise;

z^e: 1 if container i enters the yard earlier than container j and is stacked on the lower floor; 0 otherwise;

z^o: 1 if container i leaves the yard before container j and is stacked on the upper level; 0 otherwise;

z^w: 1 if the heavier container in container i and container j is stacked on the upper layer; 0 otherwise.

D. Objective Functions

The number of containers that exited first at the lower level was the smallest.

$$MinF^o = \sum_{i,j,r,t1<t2}((y - z^o) \times (t2 - t1)) \quad (1)$$

The number of containers with a high weight in the upper level was the smallest.

$$MinF^{w} = \sum_{i,j,r,t1<t2}((y - z^{w}) \times (t2 - t1)) \quad (2)$$

Minimum stacking height.

$$MinF^{h} = \sum_{i,j,r,t1<t2}(X_{i,b,r,t} \times t^{2}) \quad (3)$$

E. Constraints

$$z^{e} \geq y + E_{i,j} - 1; \forall i, j, b, r, t1 < t2 \quad (4)$$

$$z^{e} \leq y \times E_{i,j};\ \forall i, j, b, r, t1 < t2 \quad (5)$$

$$y \geq X_{i,b,r,t1} + X_{i,b,r,t2} - 1; \forall i, j, b, r, t1, t2 \quad (6)$$

$$y \leq X_{i,b,r,t1} \times X_{i,b,r,t2}; \forall i, j, b, r, t1, t \quad (7)$$

$$\sum_{i}(X_{i,b,r,t}) \geq \sum_{i}(X_{i,b,r,t+1}) \quad (8)$$

$$s_{i} = 1; \forall i \quad (9)$$

$$\sum_{i,b,r,t}(X_{i,b,r,t}) \leq b \times (r \times t - 3); \forall b, r, t \quad (10)$$

Constraints (4) and (5) indicate that z^{e} is 1 when i is placed below j and i enters the field before j, and 0 otherwise; constraints (6) and (7) indicate that the value of the decision variable y is 1 only when containers i and j are assigned to container spaces (b, r, t_1) and (b, r, t_2), respectively, and 0 otherwise; constraint (8) indicates that boxes cannot be placed in suspension; constraint (9) indicates that all hazardous materials are stored in the field area where they should be stacked; constraint (10) indicates that a buffer container space should exist within each field area.

III. Dangerous Chemicals Container Yard Bin Allocation Algorithm And Plan Optimization

A. NSGA-II Algorithm

The multi-objective functions of hazardous chemical container storage optimization are not completely co-directional functions. In most cases, optimizing one function leads to a decrease in the performance of other objective functions. Therefore, it is difficult to simultaneously optimize all objectives. Therefore, when solving the multi-objective optimization problem, the solution set obtained is optimal for one optimization objective, and may not be optimal for other optimization objectives, which causes the multi-objective function to have multiple optimal solutions. This study adopted the NSGA-II algorithm to avoid the lack of diversity of the NSGA algorithm in later stages and improved the crowding distance and crowding degree comparison operators in the algorithm. The main purpose is to maintain the diversity of the population to the best extent possible while avoiding local precocity.

Table I Comparison of traditional NSGA algorithm and improved NSGA- II shows that the improved NSGA-Ⅱ has more advantages than traditional genetic algorithms in solving multi-objective optimization problems. The steps to improve the NSGA-II algorithm are as follows.

TABLE I. Comparison of Traditional NSGA Algorithm and Improved NSGA- II

Traditional NSGA algorithm	Improved NSGA-Ⅱ algorithm
Higher computational difficulty	A fast non-dominated sorting method is proposed to reduce the computational complexity
Need to specify a shared radius	Improved crowding and crowding comparison operator to maintain the diversity of the population
No elite strategy	Introducing elite strategy, controlling elite range, and expanding sampling space

1) Initialization parameters: Generate the initial population X, the population size x_size, and the maximum number of iterations generation_size

2) Chromosome coding: This algorithm adopts the form of real number coding, as shown in Fig. 1. More specifically, there are m possible storage positions for the container after the arrival of the container, where the first n represents the distribution position and order of the container. For example, [10111 10112 10113 1012130122 10123 10131 10132 10133 10141 10142 10143], means that there are 12 locations for 1 shell in a certain area, and the storage location and order of 10 containers are [10111 10112 10113 1012130122 10123 10131 10132 10133 10141], among them, "10111" indicates that the position allocated to the stack is 1 zone, 01 shells, 1 column, and 1 floor. One chromosome corresponds to the distribution plan of the container. Under the assumption that there are n containers of the same chemical nature and m positions that can be stacked, an m-bit array needs to be generated that indicates the order in which the n containers enter the yard.

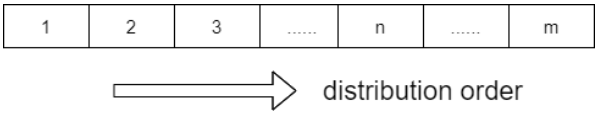


Fig. 1. Coding Scheme.

3) Fitness function: The non-dominated sorting multi-objective genetic algorithm can directly use $MinF^{o}$ $MinF^{w}$ $MinF^{h}$ as fitness functions.

4) Fast non-dominated sorting process: The solution of a multi-objective genetic algorithm is to obtain a Pareto solution set by the evolutionary approximation of the constructed genetic algorithm class. Once the fitness function is evaluated, the objective functions $MinF^{o}, MinF^{w}$, and $MinF^{h}$ are sorted by fast non-dominated solutions. The specific sorting process is illustrated in Fig. 2.

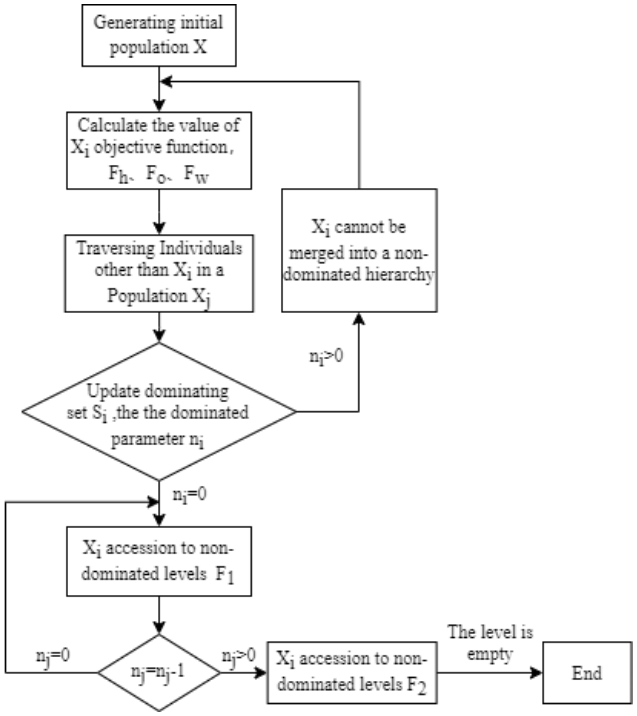


Fig. 2. Fast Non-Dominated Sorting Process.

When traversing other individuals in the population x_i, if it satisfies $F^o(x_i) > F^o(x_j)$, $F^w(x_i) > F^w(x_j)$, $F^h(x_i) > F^h(x_j)$, then it is said that individual x_i dominates individual x_j, and individual x_j is stored in the dominating set s_i. If it satisfies $F^o(x_i) < F^o(x_j)$, $F^w(x_i) < F^w(x_j)$, $F^h(x_i) < F^h(x_j)$, then it is said that individual x_j dominates individual x_i and the dominant parameter $n_i + 1$.

5) Improve the calculation method for congestion" The selection of NSGA-II will allow excellent individuals to continue to breed in iterations until the maximum population size is reached, which will easily lead to a loss of individual diversity. Ultimately, it will lead to premature convergence of the algorithm. This study has improved the algorithm to avoid obtaining the local optimal solution: first traverse the individual x_i in the non-dominated level, calculate the function value of a certain objective function, arrange it in descending order according to the function value, and set the individual congestion degree on both sides of the sequence to the maximum value that can be guaranteed to be always selected. Before calculating the crowdedness of individual x_j, first, judge whether x_j is the same as the previous individual. If they are the same, the crowning degree of individual x_j is no longer calculated, and the non-dominated level value of individual x_j is directly added to the population size. If it is not, the calculation is performed according to Equation (11). The non-dominated sorting after mixing the parent population with the offspring population produced can effectively avoid redundant individuals.

$$i_d = i_d + \frac{f_i(x_{i+1}) - f_i(x_{i-1})}{f_i^{max} - f_i^{min}} \tag{11}$$

6) Improved elite retention strategy: The elite retention strategy causes the parent and offspring to merge, and redundant individuals are prone to exist. Based on the NSGA-II algorithm, this study made some improvements to its elite retention strategy. The improved strategy is marked to judge redundant individuals and merge them into a temporary level. Finally, when the newly generated population is insufficient, the corresponding redundant individuals are removed and merged into the new population, thereby increasing the diversity of the population, as shown in Fig. 3.

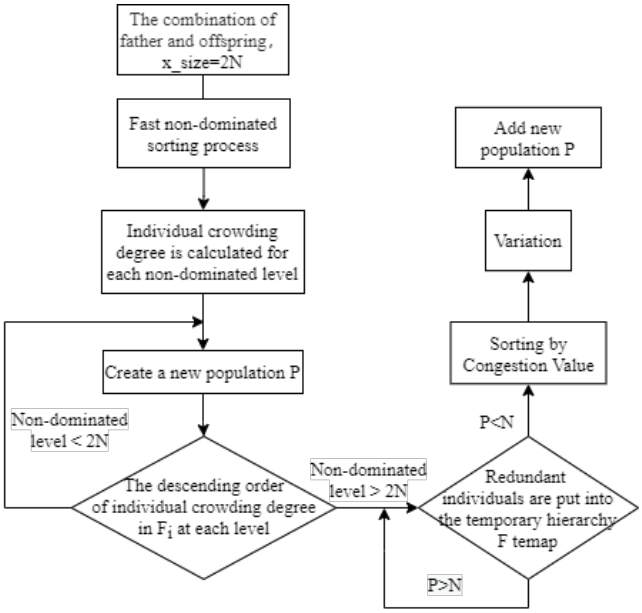


Fig. 3. Improved Elite Retention Strategy Process.

7) Genetic operation: The selection operation has adopted the roulette selection operator.

The crossover operation adopted a crossover operation to simulate a binary single-point crossover operator. The criteria were as follows:

$$\tilde{x}_{1j} = 0.5 \times \left[(1 + r_i) \cdot x_{1j}(t) + (1 - r_i)x_{2j}(t)\right]$$
$$\tilde{x}_{2j} = 0.5 \times \left[(1 + r_j) \cdot x_{1j}(t) + (1 - r_j)x_{2j}(t)\right] \tag{12}$$

$x_{i,j}$, $\tilde{x}_{i,j}$ $(i = 1,2)$ represent the j genes of the father and offspring, respectively; $r_i = \begin{cases} (2u_j)\frac{i}{\eta_c+1}, & u_j \le 0.5 \\ \left(\frac{1}{2(1-u_j)}\right)\frac{i}{\eta_c+1}, & \text{others} \end{cases}$, $u_j \in U(0,1), \eta_c > 0$ is the distribution index.

The work done in this study improved the mutation operation to determine whether the mutation is performed according to the size of the random number generated by rand (0, 1). If a mutation is needed, the gene value at one position on the individual chromosome is replaced with the gene value at another position on the chromosome.

The process set of the improved NSGA-II algorithm is shown in Fig. 4.

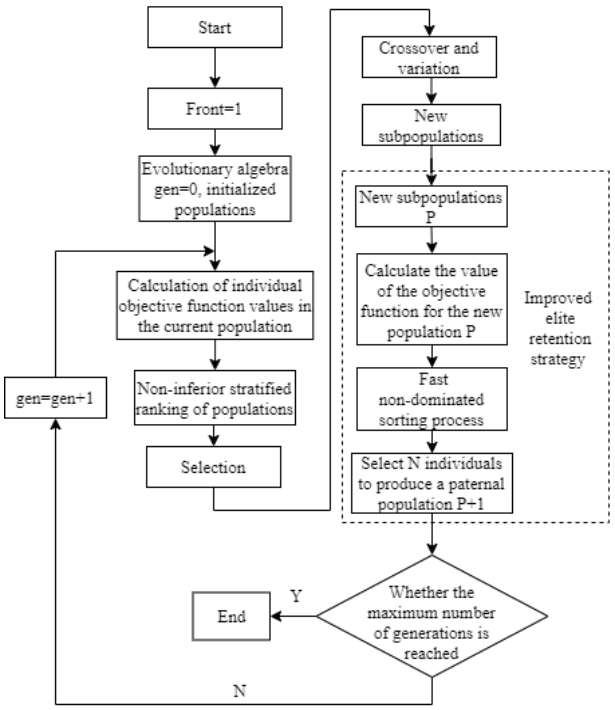


Fig. 4. Improved NSGA-II Algorithm Process.

B. Multiple Scheme Optimization

Entropy weighting is an objective method of assigning weights, which determines the weight of each indicator by the uncertainty of the information provided by the different attribute indicators. The TOPSIS method is a multi-objective decision-making analysis method that is suitable for the comparative study of multiple schemes. As a better way to avoid the subjectivity of the method, the study first used the entropy method of objective weighting to solve the weights before using the TOPSIS method to obtain Pareto optimal solution sorting. The preferred steps of the scheme are as follows:

According to the model, it can be seen that the attribute indicators affecting the decision on the stacking scheme in this paper are the order of exit, weight, and height. The entropy weight method is used to calculate the weight coefficients of these three indicators and obtain the weight matrix ω.

Use the vector normalization method to obtain the normalized weighted decision matrix.

Suppose the decision matrix $A = \{a_{ij}\}$ of the multi-attribute decision-making problem, and the standardized decision matrix $B = \{b_{ij}\}$, then

$$b_{ij} = \omega_j \cdot \frac{a_{ij}}{\sqrt{\sum_{i=1}^{m} {a_{ij}}^2}} \qquad i = 1,2,3\ldots m; j = 1,2,\ldots,n \tag{13}$$

Apply the weighted distance to construct the Euclidean distance between the target solution and the ideal solution and the negative ideal solution.

$$d_j = \sqrt{\sum_j^n (x_{ij} - x_j)^2} \qquad i = 1,2,3\ldots m \tag{14}$$

Calculate the comprehensive evaluation index of each plan and rank the superiority and inferiority of the plan according to the value C_i in descending order. Get the optimal stacking solution.

$$C_i = \frac{d_i^-}{d_i^+ + d_i^-} \qquad i = 1,2,3\ldots m \tag{15}$$

Based on the above analysis, the process settings of the improved NSGA-Ⅱ algorithm are shown in Fig. 5.

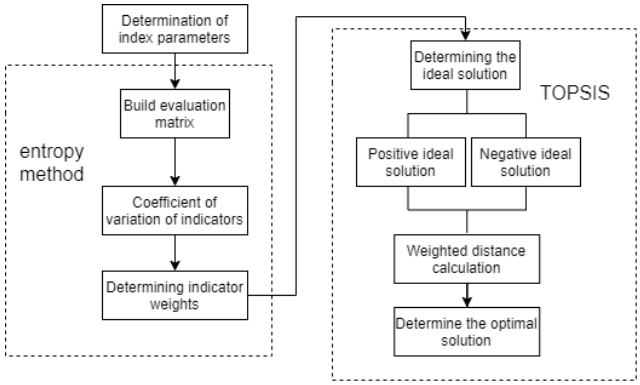


Fig. 5. Entropy TOPSIS Method Calculation Process.

IV. Case Analysis

A. Known Conditions

Consider a container area of a hazardous chemical container yard in a port as an example. The considered container area has 14 shell positions, each of which has three rows, and the maximum stacking height is three layers. The study randomly selected 50 containers of hazardous chemicals to be processed from 0:00 to 1:00 on June 2, 2020. The types of substances in the containers include category 6.1 (toxic substances), category 8 (corrosive substances), and category 9 (miscellaneous hazards). After the selected containers are classified according to the existing isolation rules of the storage yard, the 43 selected containers are all stacked in the same content area, and stacking is allowed for up to three layers. The initial number of layers of each shell in the container area is shown in Fig. 6.

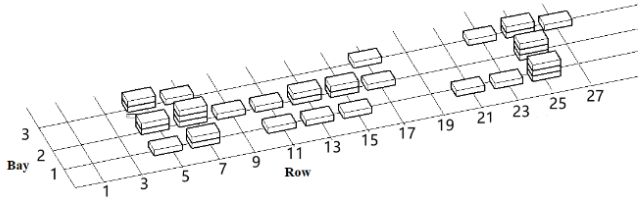


Fig. 6. The Initial Number of Stacking Levels in Each bay of the Box Area

The information of 43 containers to be processed is shown in Table II.

B. Optimization Results and Analysis

1) Improve the calculation method for congestion: The algorithm of the stack optimization model established in Section III is used, and the algorithm settings are as follows:

Population size $x\_size = 100$, maximum iteration $generation\_size = 500$, crossover probability $p = 0.8$; the mutation probability $q = 0.02$.

MATLAB was used to compile the code and run the program, and the iteration was terminated 500 times. The fitness of the iterative process changes, as shown in Fig. 7.

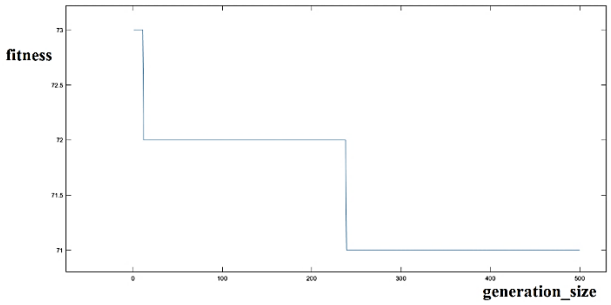


Fig. 7. Improved NSGA-II Algorithm Convergence Process.

As the number of iterations increases, the fitness value of the chromosome eventually converges, and a good convergence effect is achieved. After the program runs, 92 sets of Pareto solutions that simultaneously satisfy the requirements were obtained (Table III).

Based on the 92 Pareto solutions obtained, a 92 × 3 decision matrix A was constructed, and the attribute weights of the order of appearance, weight, and height were obtained using the entropy method in the following way: w_1=0.4685, w_2 =0.2954, w_3 =0.2360. It is evident that the order of appearance has the greatest impact on the container storage plan.

Based on the 92 Pareto solutions obtained, this study constructs a 92 × 3 decision matrix A (results in the C_i column in Table II). According to the TOPSIS method evaluation criteria, the solution with the highest score is selected as the final satisfactory solution; that is, the 34th solution set in Table III is the heap, and the optimal solution is stored. The stacking plan for the solution set is shown in Fig. 8.

TABLE II. COMPARISON OF TRADITIONAL NSGA ALGORITHM AND IMPROVED NSGA- II

Number	Shipping company	Weight	Departure time	Category	Number	Shipping company	Weight	Departure time	Category
1	HLC	28384	2020/6/2 15:04	8	23	POL	22160	2020/6/3 17:06	8
2	SNL	28314	2020/6/3 0:04	8	24	CMA	22200	2020/6/3 17:06	8
3	MSC	24628	2020/6/3 11:28	8	25	WHL	22960	2020/6/3 17:16	8
4	MSC	24305	2020/6/3 11:29	8	26	HMM	22940	2020/6/3 17:16	8
5	MSC	24305	2020/6/3 11:29	8	27	HMM	23050	2020/6/3 17:31	6.1
6	SIT	27285	2020/6/3 11:37	8	28	HMM	27330	2020/6/3 17:36	9
7	UAS	27285	2020/6/3 12:15	8	29	OOL	27330	2020/6/3 17:36	9
8	SNL	27285	2020/6/3 12:16	8	30	MSC	23128	2020/6/3 17:59	6.1
9	MSC	27285	2020/6/3 12:18	8	31	WHL	22245	2020/6/3 17:59	9
10	MSC	27285	2020/6/3 12:18	8	32	WHL	22260	2020/6/3 18:01	9
11	SCL	25280	2020/6/3 12:22	6.1	33	CMA	22220	2020/6/3 18:01	3
12	SCL	29388	2020/6/3 13:39	8	34	NYK	22300	2020/6/3 18:02	8
13	MKL	29328	2020/6/3 13:39	8	35	SNL	14498	2020/6/3 18:03	8
14	MSC	2716	2020/6/3 13:50	8	36	APL	27300	2020/6/4 1:51	6.1
15	CNC	19178	2020/6/3 16:42	8	37	KMT	29230	2020/6/5 11:46	8
16	MSC	19440	2020/6/3 16:42	8	38	KMT	29230	2020/6/5 11:47	8
17	POL	22360	2020/6/3 16:48	8	39	KMT	29180	2020/6/5 11:48	8
18	POL	22360	2020/6/3 16:48	8	40	WHL	29230	2020/6/5 11:48	8
19	POL	19420	2020/6/3 16:50	8	41	WHL	29200	2020/6/5 11:50	8
20	POL	19420	2020/6/3 16:50	8	42	SNL	29180	2020/6/5 11:50	8
21	POL	19440	2020/6/3 16:51	8	43	HLC	29230	2020/6/5 11:52	8
22	POL	19460	2020/6/3 16:51	8					

TABLE III. COMPARISON OF TRADITIONAL NSGA ALGORITHM AND IMPROVED NSGA-II

Number	F^h	F^o	F^w	C_i	Number	F^h	F^o	F^w	C_i	Number	F^h	F^o	F^w	C_i
1	71	7	8	0.0104	32	72	6	9	0.0102	63	71	11	8	0.0157
2	71	5	7	0.0061	33	72	5	9	0.009	64	71	9	10	0.0164

3	72	7	10	0.0129	34	71	10	12	0.0202	65	71	9	7	0.0122
4	72	6	8	0.0089	35	72	7	9	0.0117	66	72	5	10	0.0102
5	72	9	9	0.0151	36	71	7	10	0.0129	67	71	9	8	0.0136
6	71	6	10	0.0114	37	71	6	7	0.0074	68	71	4	9	0.0081
7	71	7	5	0.0068	38	72	7	7	0.0091	69	71	4	7	0.0053
8	71	11	10	0.0188	39	71	6	5	0.0049	70	72	9	11	0.0176
9	72	5	8	0.0076	40	71	5	8	0.0076	71	71	10	7	0.0135
10	71	8	9	0.0134	41	72	8	11	0.0156	72	72	9	10	0.0164
11	72	6	6	0.006	42	72	8	8	0.0121	73	71	7	11	0.0139
12	71	8	8	0.0121	43	72	8	6	0.0095	74	73	4	5	0.002
13	71	10	10	0.018	44	71	6	9	0.0102	75	73	6	7	0.0074
14	72	10	9	0.0164	45	73	7	8	0.0104	76	71	5	4	0.0022
15	72	7	5	0.0068	46	71	10	9	0.0164	77	71	4	6	0.0037
16	71	6	6	0.006	47	71	5	10	0.0102	78	71	8	5	0.0086
17	72	7	6	0.0078	48	71	8	7	0.0107	79	72	11	7	0.0143
18	71	7	6	0.0078	49	71	5	9	0.009	80	71	11	7	0.0143
19	71	6	8	0.0089	50	72	7	11	0.0139	81	73	8	7	0.0107
20	72	6	7	0.0074	51	72	8	4	0.0079	82	72	9	7	0.0123
21	72	7	8	0.0104	52	71	7	7	0.0091	83	71	11	6	0.0132
22	71	8	11	0.0156	53	71	6	4	0.0043	84	72	8	7	0.0107
23	71	10	8	0.0149	54	72	6	5	0.0049	85	72	9	9	0.0151
24	71	9	9	0.015	55	73	7	7	0.0091	86	72	10	6	0.0123
25	71	8	4	0.0079	56	72	6	10	0.0114	87	71	9	11	0.0175
26	71	7	12	0.0146	57	71	5	5	0.003	88	72	5	5	0.003
27	72	5	6	0.0045	58	71	5	6	0.0045	89	71	10	6	0.0123
28	71	8	6	0.0095	59	71	7	9	0.0117	90	71	11	9	0.0171
29	72	5	7	0.0061	60	71	4	8	0.0068	91	72	9	8	0.0136
30	72	8	10	0.0147	61	72	8	9	0.0134	92	72	10	8	0.0149
31	71	8	10	0.0146	62	72	10	10	0.018					

TABLE IV. THE OPTIMAL PARETO SOLUTION SET OF TRADITIONAL NSGA ALGORITHM

F^h	F^o	F^w	F^h	F^o	F^w	F^h	F^o	F^w	F^h	F^o	F^w	F^h	F^o	F^w	F^h	F^o	F^w	F^h	F^o	F^w
77	3	7	78	4	6	78	3	8	79	0	7	79	2	4	77	2	6	76	3	7
77	4	5	77	4	6	72	5	11	79	2	5	74	5	6	77	5	7	76	6	10
77	3	5	77	2	4	76	2	7	78	6	7	79	2	9	77	5	6	77	2	9
78	2	5	77	0	4	76	2	6	76	7	4	79	4	7	77	3	7	76	5	10
77	3	3	78	3	4	78	3	2	78	2	9	78	4	4	77	1	6	76	4	10
78	1	8	78	5	6	78	4	10	78	2	4	77	2	8	77	4	4	78	5	3
77	1	4	78	4	8	77	6	10	76	4	9	76	4	7	77	2	3	76	4	11
77	3	6	78	5	7	76	6	5	76	5	4	76	8	7	78	6	6	79	3	4
77	4	7	78	4	5	78	1	9	75	6	10	79	1	4	78	3	5	76	5	7
71	5	11	77	6	5	75	5	10	78	4	7	77	5	8	77	6	6	78	2	8
77	4	3	79	4	6	78	1	7	79	8	6	77	7	9	77	7	6	75	5	7
77	3	4	76	3	5	75	7	9	76	5	6	77	4	10	79	6	6	74	4	11

78	3	9	76	4	6	79	4	5	73	5	10	76	6	6	77	1	7	75	6	6
77	5	5	76	5	5	79	5	7	78	3	10	76	3	10	78	3	7	77	3	9
79	4	4	77	1	3	78	5	5	78	5	9	76	4	8	76	3	6	74	3	6
73	4	11	77	2	7	77	4	9	78	7	5	79	1	8	77	6	7	78	5	10
77	5	4	78	6	3	74	3	11	78	6	5	79	2	6	79	2	7	77	6	4
78	3	6	78	3	3	78	2	7	77	7	5	77	1	2	78	2	6	74	5	12
77	2	2	78	4	3	75	7	7	72	5	12	79	4	3	78	5	4	74	6	8
77	2	5	77	4	8	76	7	10	79	5	6	78	6	8	79	1	9	77	0	7
78	5	8	77	5	3	77	3	8	77	3	2	78	8	6	79	3	8	76	6	4
79	6	7	77	1	5	79	3	7	76	4	4	74	5	9	73	9	10			

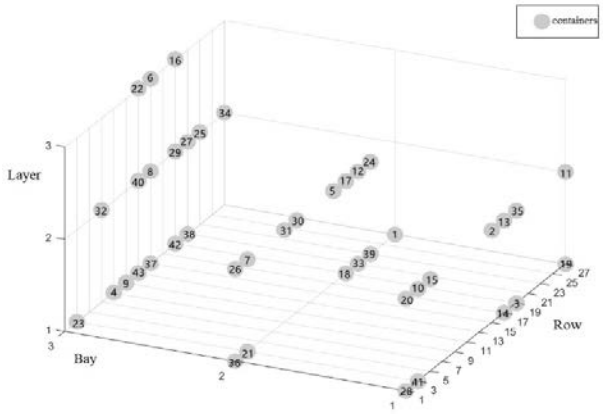


Fig. 8. Box Allocation Result.

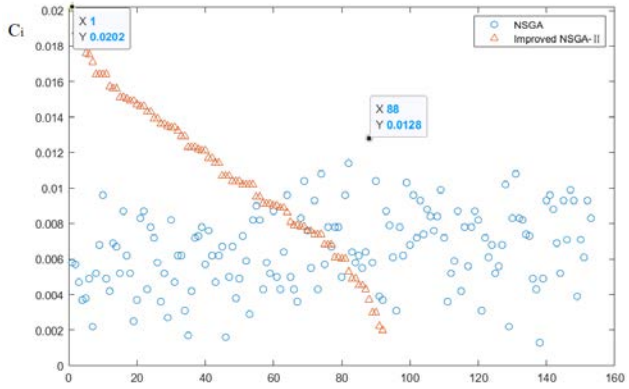


Fig. 9. Comparison between Traditional NSGA Algorithm and Improved NSGA-Ⅱ.

In practical engineering applications, decision-makers can choose other non-inferior solutions given by the algorithm according to the actual situation. For instance, when weight is a prominent consideration, the scheme with the smallest F^w can be selected, and when height balance is a prominent consideration, the scheme with the smallest F^h can be selected.

2) Algorithm performance analysis: In this study, we used the same container data and crossover and mutation probabilities to verify the effectiveness of the improved algorithm and solve it with the traditional NSGA algorithm. In total, 153 Pareto solution sets were obtained. (Table IV).

The results show:

a) Under identical conditions, for 500 iterations, the comprehensive scores of the traditional NSGA algorithm and the improved NSGA-II algorithm are shown in Fig. 9. It is evident that the highest weighted comprehensive score of the objective function obtained by the traditional NSGA algorithm is lower than that of the improved NSGA-II algorithm, which indicates that the solution obtained by the improved NSGA-II algorithm is closer to the optimal solution.

b) After 500 iterations, the operational time of the traditional NSGA algorithm was 68.8 s, and the time of the improved NSGA -II algorithm was 9.86 s, which is a reduction by 85.7%, indicating that the time consumed by the improved NSGA -II algorithm is considerably shorter.

V. Conclusion

The improvement of operational efficiency and the optimization of stacking safety have always been critical issues in the management of yards containing dangerous goods containers at ports. Combined with the needs of the application level, a multi-objective optimization model was constructed and the algorithm was optimized. The following conclusions can be drawn:

1) Combined with the operation process of the dangerous goods container yard and the technical requirements of the storage, a multi-objective optimization model is established, which realizes the dual consideration of safety and efficiency, and provides the theoretical basis and technical support for the safe and efficient operation of the dangerous goods container yard.

2) The NSGA-II algorithm was introduced to solve the optimization model and the algorithm was tested and improved. Compared with the traditional NSGA algorithm, the running time of the improved algorithm was shortened by 85.7%, which not only improved the efficiency of the algorithm but also enriched the diversity of understanding. The sorting algorithm improves the pertinence of the solution.

3) The container location allocation in this study focuses on storage optimization within the same container area. The coordinated storage problem of multiple container areas for

multi-category containers will be further explored in the future study.

ACKNOWLEDGMENT

This study was supported by Natural Science Foundation of Shan-dong Province (ZR2020MB126, ZR2019MG023), People's Livelihood Science and Technology Project of Qingdao (19-6-1-84-nsh), National Natural Science Foundation of China (71403293, 51104174), Shandong Province Graduate Education Quality Improvement Plan Project (SDYAL19034).

REFERENCES

[1] C. Zhang, M. Zhong, and L. Miao, "Location assignments for outbound containers in container terminals," Journal of Tsinghua University (Science and Technology), vol. 55, no. 10, pp. 1150-1156, 2015, doi: 10.16511/j.cnki.qhdxxb.2015.22.001.

[2] D. Ambrosino and H. Xie, 'Optimization approaches for defining storage strategies in maritime container terminals', Soft Comput, pp. 1–13, Feb. 2022, doi: 10.1007/s00500-022-06769-7.

[3] M. D. S. De Marcos, S. Toulouse, and R. W. Calvo, 'A new effective unified model for solving the Pre-marshalling and Block Relocation Problems', European Journal of Operational Research, vol. 271, 2018, doi: 10.1016/j.ejor.2018.05.004.

[4] Q. Q. Shao, Q. Xu, Z. Bian, and Z. H. Jin, "Stockpiling operating optimization for yard crane with containers delivery time uncertainty," Xitong Gongcheng Lilun yu Shijian/System Engineering Theory and Practice, vol. 35, no. 2, pp. 394-405, 2015, doi: JournalArticle/5b3b7f46c095d70f0078dc8e.

[5] C. J. Liu and Z. H. Hu, "Multi-objective optimization model for storage location allocation of outbound containers at container yard," Journal of Dalian University of Technology, vol. 55, no. 6, pp. 589-596, 2015, doi: 10.7511/dllgxb201506005.

[6] Y. He, A. Wang, and H. Su, 'The impact of incomplete vessel arrival information on container stacking', International Journal of Production Research, vol. 58, no. 22, pp. 6934–6948, Nov. 2020, doi: 10.1080/00207543.2019.1686188.

[7] H. Yu, J. Ning, Y. Wang, J. He, and C. Tan, 'Flexible yard management in container terminals for uncertain retrieving sequence', Ocean & Coastal Management, vol. 212, p. 105794, Oct. 2021, doi: 10.1016/j.ocecoaman.2021.105794.

[8] D. Ku and T. S. Arthanari, 'Container relocation problem with time windows for container departure', European Journal of Operational Research, vol. 252, no. 3, pp. 1031–1039, Aug. 2016, doi: 10.1016/j.ejor.2016.01.055.

[9] H. M. Fan, X. Yao, and M. Z. Ma, "Storage space allocation based on regional workload balance planning of multiple yard cranes in container terminal yard," Control and Decision, vol. 31, no. 9, pp. 1603-1608, 2016, doi: 10.13195/j.kzyjc.2015.1101.

[10] H. X. Zheng, B. L. Liu, H. B. Kuang, and X. Yan, "Multi-yard Cranes Scheduling Optimization of Export Container Yard Considering Real-time Pre-marshaling," Chinese Journal of Management Science, vol. 26, no. 9, pp. 85-96, 2018, doi: 10.16381/j.cnki.issn1003-207x.2018.09.009.

[11] X. Zhou, Q. Miu, Y. Shen, and Y. Li, "Storage location allocation model for dangerous goods containers at container yard," Containerization, vol. 30, no. 01, pp. 17-19, 2019, doi: 10.13195/j.kzyjc.2015.1101.

[12] H. Zhu, M. Ji, W. Guo, Q. Wang, and Y. Yang, 'Mathematical formulation and heuristic algorithm for the block relocation and loading problem', Naval Research Logistics, vol. 66, no. 4, pp. 333-351, Jun. 2019, doi: 10.1002/nav.21843.

[13] X. Mengjue, Z. Ning, and M. Weijian, 'Storage Allocation in Automated Container Terminals: The Upper Level', Pol. Marit. Res., vol. 23, pp. 160-174, 2016, doi: 10.1515/pomr-2016-0061.

[14] Y. Zhao, Q. Xue, and X. Zhang, 'Stochastic Empty Container Repositioning Problem with CO2 Emission Considerations for an Intermodal Transportation System', Sustainability, vol. 10, no. 11, Art. no. 11, Nov. 2018, doi: 10.3390/su10114211.

[15] WangTiantian, MaHong, XuZhou, and XiaJun, 'A new dynamic shape adjustment and placement algorithm for 3D yard allocation problem with time dimension', Computers & Operations Research, vol. 138, p. 105585, Feb. 2022, doi: 10.1016/j.cor.2021.105585.

[16] C. Liang, L. Yingbo, D. Wang, L. Center, and S. M. University, "RESEARCH ON THE YARD CRANE SCHEDULING PROBLEM BASED ON ROLLING WINDOW STRATEGY," Computer Applications and Software, vol. 335, no. 1, pp. 72-76+127, 2018, doi: CNKI:SUN:JYRJ.0.2018-01-013.

[17] T. Kim and K. R. Ryu, 'Deriving Situation-Adaptive Policy for Container Stacking in an Automated Container Terminal', Applied Sciences, vol. 12, no. 8, Art. no. 8, Jan. 2022, doi: 10.3390/app12083892.

[18] X. Tang, J. G. Jin, and X. Shi, 'Stockyard storage space allocation in large iron ore terminals', Computers & Industrial Engineering, vol. 164, p. 107911, Feb. 2022, doi: 10.1016/j.cie.2021.107911.

[19] Y. E. Jun and L. Chen, "Mathematical Model and Genetic Algorithm for Multiple Yard Cranes to Load Outbound Containers in a Maritime Terminal," Industrial Engineering & Management, vol. 22, no. 4, pp. 100-106+114, 2017, doi: 10.19495/j.cnki.1007-5429.2017.04.014.

Word by Word Labelling of Romanized Sindhi Text by using Online Python Tool

Irum Naz Sodhar[1]
Post-Doctoral Fellow, Department of Computer Science, Kulliyyah (Faculty) of Information and Communication Technology, International Islamic University Malaysia

Abdul Hafeez Buller[2]
Post-Doctoral Fellow, Department of Civil Engineering, Kulliyyah (Faculty) of Engineering, International Islamic University Malaysia

Suriani Sulaiman[3]
Assistant Professor, Department of Computer Science, Kulliyyah (Faculty) of Information and Communication Technology, International Islamic University Malaysia

Anam Naz Sodhar[4]
Postgraduate Student, Quaid-e-awam University of Engineering, Science & Technology, Nawabshah, Sindh, Pakistan

***Abstract*—Sindhi is one of the most ancient languages in the world and it has its own written and spoken scripts. After the rigorous study it was found that a lot of research work has been done in different languages, but word by word labelling of Sindhi language had not been done yet. In this research study, word labelling was done on 100 sentences of Romanized Sindhi texts using Python online tool. The dataset was collected from different sources which include Sindhi newspaper, blogs and social media webpages. From this dataset, a rule-based model has been applied for the Parts-of-Speech (POS) tagging of the Romanized Sindhi sentences. A total of 624 words of Romanized Sindhi texts were tested and successfully tagged by the SindhiNLP tool in which 482 words were tagged as nouns and pronouns, 92 words tagged as verbs and 50 words tagged as determinants.**

Keywords—Romanized sindhi; word labelling; rule-based model; POS tagging; SindhiNLP tool

I. Introduction

Sindhi is one of the most ancient languages in the world which has its own script in written and spoken forms [1-3]. Communication technologies are increasing day-by-day for different purposes, while different applications and software are used for daily communications such as WhatsApp, Facebook, Twitter, Telegram and Instagram [4-5]. In the community that uses Sindhi as their main language, Romanized Sindhi texts are used in daily communication especially in writing text messages on mobile phones, WhatsApp and other social media platforms [6].

Natural Language Processing has a vital role in the field of machine learning. This field provides language processing tasks such as of Parts–of–Speech tagging, tokenization of text (i.e., words, sentences, and paragraph) to the users [7-8]. In this research study, 100 sentences of Romanized Sindhi texts were labelled. The word labelling process which consists of two natural language processing tasks which is tokenization and POS tagging was performed using an online SindhiNLP tool [9]. Before performing the two tasks, a rule-based model has been applied for the POS tagging of the sentences to improve the accuracy of the POS tags [10-12].

After the review of the literature it was observed that a lot of vacuum is still available for the Sindhi language. This research study presents the word by word labelling of Sindhi language after Romanization.

II. Methodology for Labelling of Romanized Sindhi

The procedure for labelling of the Sindhi Romanized text has been divided into various stages as shown in Fig. 1. The first phase involves the data collection process from different sources of Sindhi scripts, the second stage is the conversion of Sindhi scripts into Romanian scripts (i.e., Romanization), the third stage identifies the issues in word labelling after applying the rule-based model and the final task is to do a thorough analysis on the results produced [13].

A. Dataset of Sindhi Text

Sindhi language is one of the oldest, historical and most commonly used languages in the world. Sindhi language is more difficult than other languages due to the difficulty in reading, writing and understanding the scripts [13-14]. Sindhi language is spoken by the people in the province of Sindh which is the second largest populated province of Pakistan. Sindhi is the official language of the Sindh province in which almost 15% of the population use Sindhi as their mother tongue [14-15]. As Sindhi language is mostly used in Sindh-Pakistan, the data for this research study was collected within the province of Sindh. Data was collected from different sources (Sindhi newspaper, blogs, and social media webpages) which provided the rules and guidelines of Romanized Sindhi for text communication.

B. Sindhi Alphabet

Sindhi language has its own script and written style like other languages (Arabic, Urdu, and English) [16]. In Sindhi script there are 52 alphabetical letters for writing and speaking purposes and presented in Fig. 2. Sindhi language has one of the largest numbers of alphabetical letters as compared to other languages. Similar to Arabic and Urdu scripts, the Sindhi script is written from right to left with a total of 52 alphabets [17].

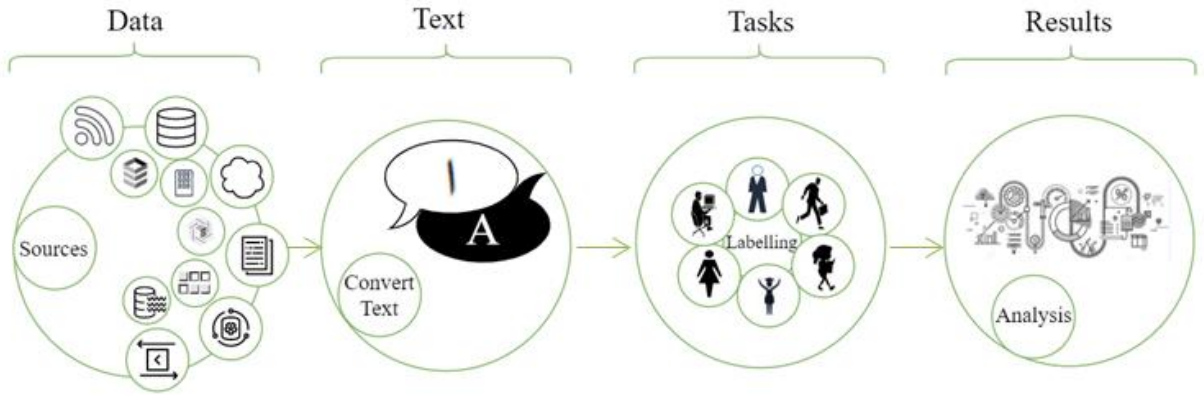


Fig. 1. Methodology of Labelling of Sindhi Text.

Sindhi Alphabet									
ٿ	ٽ	ٺ	ٹ	ت	ٻ	ب	ب	ا	Sindhi alphabet
s	th	t	th	t	bh	b	b	a, u	Roman
ڇ	ڃ	ڇ	ڇ	ج	ڄ	ج	ج	پ	Sindhi alphabet
kh	h	ch	ch	j	jh	j	j	p	Roman
ز	ڑ	ر	ذ	ڊ	ڊ	ڈ	ڈ	د	Sindhi alphabet
z	r	r	z	dh	d	d	dh	d	Roman
ف	غ	ع	ظ	ط	ض	ص	ش	س	Sindhi alphabet
f	ga	a	z	t	z	s	sh	c, s	Roman
ل	ڳ	ڳھ	ڳ	ڳ	ک	ڪ	ق	ڦ	Sindhi alphabet
l	g	gh	ga	g	kh	k	q	ph	Roman
		ي	ء	ھ	و	ڻ	ن	م	Sindhi alphabet
		y, e	a	h	o, w, v	n	n	m	Roman

Fig. 2. Sindhi-Roman Alphabet [17].

C. Romanization of Sindhi Text

In this research study, 100 sentences were used for the word labelling of Sindhi texts. After the collection of Sindhi sentences for the data set for this research study, the collected dataset was converted from Sindhi scripts into Romanized Sindhi text by using rules for Romanization of Sindhi text. Romanization of Sindhi text was successfully done following the rules for Romanized Sindhi text.

III. Pre-processing of Romanized Sindhi

Pre-processing is the basic components of NLP to filter the raw the data to useful and remove unnecessary data from the text. The pre-processing step consists two steps first is performing tokenization and second one assigning tag on each token [10].

A. Tokenization of Romanized Sindhi Text

The tokenization of Romanized Sindhi text has been done using the online SindhiNLP Python tool [9]. The Romanized texts were prepared following the rules of Sindhi on 100 sentences. The statistical information after the tokenization process of the Sindhi text is shown in Table I. This table consists of five different columns which are: total number of sentences, total number of words, total number of characters (with space), total number of character (without space) and total number of word tokens. In this table, two types of sentences were used: sentences from Sindhi text and sentences from Romanized Sindhi text. A total of 652 words, 2,816 characters with space and 2,262 characters without space were extracted as shown in below Table I.

TABLE I. Statistical Study Data of Token

Description	*Total number of sentences*	*Total number of words*	*Total number of characters (with space)*	*Total number of characters (without space)*	*Total number of word tokens*
Sentences in Sindhi scripts	100	652	2816	2262	---
Sentences in Romanized Sindhi	100	624	3275	2740	624

B. Parts-of-Speech Tagging

The POS tagging task for Sindhi was designed such that the whole process was divided into a few steps. The first step involved the pre-processing of the Romanized Sindhi sentences. Subsequently, the ruled-based model of Sindhi was applied for the Romanization process as described in Table II. This Romanized Sindhi text was then used as input to the SindhiNLP tool, after the input Romanized Sindhi text, the text was pre-processed using the online SindhiNLP Python tool [9] in which the sentences were split into words (i.e, tokenization). Next, the Match step was performed which was also sub-divided into two categories: Assigned Tag and Incorrect Tag. If the tag was incorrectly assigned, we apply the rule-based model and repeat the process again.

C. Algorithm for POS Tagging of Romanized Sindhi Text

The algorithm for the Parts of speech tagging of Romanized Sindhi text was designed before the start of the research work. The algorithm used was based on the ten steps described below. The same step applies following the algorithm for every new input data of Romanized Sindhi text.

Step 0 Start

Step 1 Take input sentence

Step 2 Split text → words

Step 3 Repeat steps 2→7 when ⩾ get appropriate output

Step 4 If word is matched, continue to assign tag separately, word by word

Step 5 If same tag is assigned to multiple words, apply rules for words and assign one tag for each word

Step 6 If one tag is assigned to one word, display the word with tag

Step 7 Else, select one or more morphological rules and apply to words to extract word with appropriate tag.

Step 8 Display as output the tagged words

Step 9 Apply rules for new words when entered

Step 10 End

D. Rule-Based Model for Labelling of Romanized Sindhi Text

The rule-based model used in the word labelling of Sindhi text is a supervised machine learning model or hybrid model. This model combines the use of online and manual approach. This type of model is commonly used to create rules for language analysis and is a popular NLP technique to perform different tasks on different languages as it is easier to understand while the results are based on ground truth values [19-20]. Fig. 3 illustrates that the S1, S2, S3 until Sn are input sentences while R1, R2, R3 until R10 are the rules. These rules are applied on the Sindhi sentences to get the appropriate output, Y. The rules for Romanized Sindhi texts are described in Table II.

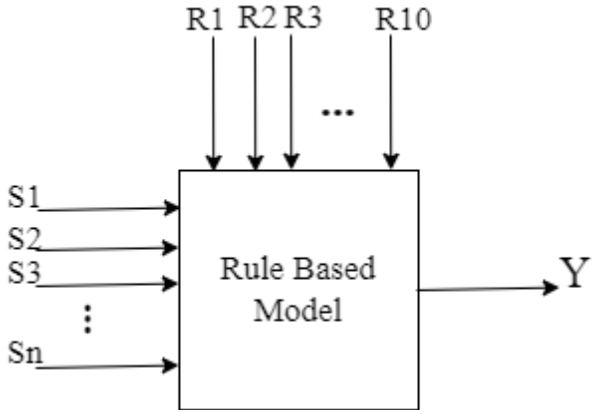


Fig. 3. Rule-Based Inputs and Output for Sindhi POS Tagging

There are ten rules that have been created for the word labelling of Romanized Sindhi texts [21-22]. Rule 1 describes the structure of a sentence and the restructuring of an input sentence by applying the SVO structure (Subject + Verb + Object) [18]. Rule 2 is used to define the prefixes of Sindhi sentences (i.e., ma, mounkhe, huwa, manhon, na, wanu sijh, cha, eho, kethe, Ali, Sara) as starting words and refers to nouns. Rule 3 describes the prefix that appears in sentences (i.e., he) as an initial word which is considered as pronouns. Rule 4 is used for the words that appear at the beginning of input sentences (i.e., Ma, Mounkhe, Huwa, Manhon, Na, Wanu sijh, cha, eho, kethe, Ali, Sara, he, etc.), considered as nouns as well as pronouns. Rule 5 describes the words that appear in the middle of an input sentence (i.e., Sadyo, Parhyo, maryo, likhyo, budho, khedan) known as the verb class. Rule 6 is used when the infix letters (i.e., a, d, e and o) appear in between words in a sentence which refers to a verb class. Rule 7 is used for postfix letters (i.e., e, o, n, i, u), if they appear in the middle of a word in a sentence which refers to a verb class. Rule 8 is used for the postfix letters (i.e., d, e, h, o, and y) if they appear at the end of the final word in a sentence, which belongs to a noun class. Rule 9 applies when the part-of-speech tagger fails to identify when the input sentences are interrogative. Rule 10 is used when the parts-of-speech tagging is performed on sentences with negation (without subject in the sentence), otherwise it was not identified. The rules used for Romanized Sindhi Text help in performing POS tagging on the SindhiNLP tool [9] to produce a more accurate part-of-speech.

TABLE II. RULES FOR ROMANIZED SINDHI TEXT FOR POS USING THE TEMPLATE

R #	Rule Description	Related Examples
1	Sentence structure should be built by applying the SVO (Subject +Verb +Object) structure.	You are teacher ⇩ ⇩ ⇩ Subject Verb Object ⇩ ⇩ ⇩ Tou ahen teacher
2	Prefixes (Ma, Mounkhe, Huwa, Manhon, Na, Wanu sijh, cha, eho, kethe, Ali, Sara etc.) in sentences as starting words, refers to noun class.	I am a Student ⇨ Ma/NNP ahyan/VBD shagrid/JJ ⇨ مان آهيان شاگرد ⇨
3	Prefix (he) in sentences as starting words, refers to pronoun class.	He is intelligent He/PRP ahy/VBD hoshar/NN ⇨ هي آهي هوشيار
4	Prefixes (Ma, Mounkhe, Huwa, Manhon, Na, Wanu sijh, cha, eho, kethe, Ali, Sara, he etc.) in sentences as starting words, refers to noun as well as pronoun class.	I play game ⇨ Ma/NNP khedan/VBD rand/NN ⇨ مان کيڈان راند
5	Infixes (Sadyo, Parhyo, maryo, likhyo, budho, khedan etc.)	I wrote article ⇨ Ma/ NNP likhyo/VBD article/NN ⇨ مان لکيو آرٹيڪل

	that appear in the middle of sentences, known as verb class.	
6	Infixes (a, d, e, o) that appear in the middle of the words in sentences refers to verb class.	I am happy Ma/NNP ahyan/VBDkush/JJ ⇨ مان آهيان خوش
7	Postfixes (e, o, n, i, u) that appear in the middle of the words in a sentence refers to verb class.	I learn Sindhi ⇨ Ma/NNP sikhan/VBD thi/NN Sindhi/NNP ⇨ مان سکان ٿي سنڌي
8	Postfixes (d, e, h, o, and y) that appears at the end of the last words in a sentence refers to noun class.	You are teacher ⇨ Tou/NNP ahen/VBD ustad/NN ⇨ تون آهين استاد
9	Parts of speech not identified when sentence is interrogative.	Do I like banana? Kayan/NN thi/NN ma/NN pasand kela/NN? ⇨ کيان ٿي مان پسند کيلا
10	Parts of speech perform on sentences with negation (without the subject in the sentence), otherwise not identified.	Do not forget ⇨ Na/NNPwesaryo/ VBD ⇨ نه وساريو Negative Verb

IV. Word by Word Labelling of Romanized Sindhi

Word labelling of Romanized Sindhi Text was performed using the free online SindhiNLP Python tool [9]. Word Labelling of Romanized Sindhi text has been performed after completing the two pre-processing tasks for Sindhi Romanized text: the tokenization and part-of-speech tagging tasks as shown in Table III.

TABLE III. Word by Word Labelling of Romanized Sindhi Text (Examples)

#	Sindhi Sentence	English Sentence	Romanized Sindhi	Word Tokens	Word Labelling
01	اهي پلي هاڻي ڪم ڪن	They better work now	Ehe kamu kan bhale hanne	Ehe, kamu, kan, bhale, hanne	Tagged Text: Ehe/NNP kamu/VBD kan/NN bhale/NN hanne/NN
02	اسين هاڻي پلي آرام ڪريون	We should rest now	Aseen kryon bhale hanne aram	Aseen, kryon, bhale, hanne, aram	Tagged Text: Aseen/NNP kryon/VBD bhale/NN hanne/NN aram/NN
03	هوءَ هڪ ڊاڪٽر هُئي	She was a doctor	Huoa hue hek doctor	Huoa, hue, hek, doctor	Tagged Text: Huoa/NNP hue/VBP hek/NN doctor/NN
04	مان ڪراچيءَ ۾ هيس	I was in karachi	Maa'n huoas Karachi maen	Maa'n, huoas, Karachi, maen	Tagged Text: Maan/NNP huoas/VBD Karachi/NNP maen/NN
05	توهان هڪ خوبصورت ڇوڪرا هُو	You are a handsome boy	Tawhan Huoao hek khubhsorat chokra	Tawhan, Huoao, hek, khubhsorat, chokra	Tagged Text: Tawhan/NNP Huoao/NNP hek/NN khubhsorat/NN chokra/NN
06	هوءَ هڪ موهيندڙ ڇوڪري هُئي	She was an attractive girl	Huoa hui hek mohendar chokri	Huoa, hui, hek, mohendar, chokri	Tagged Text: Huoa/NNP hui/NN hek/NN mohendar/NN chokri/NN
07	اهو ڏکوئيندڙ هو	It was painful	Eho dukhoindar ho	Eho, dukhoindar, ho	Tagged Text: Eho/NNP dukhoindar/NN ho/WP
08	اسين هتي آفيس ۾ هئاسين	We were here in the office	Aseen huoaseen hite office maen	Aseen, huoaseen, hite, office, maen	Tagged Text: Aseen/NNP huoaseen/VBD hite/JJ office/NN maen/NN
09	هي پهرين سال واري ڪلاس ۾ هو	He was in the first year class	He ho pehreyen saal ware class maen	He, ho, pehreyen, saal, ware, class, maen	Tagged Text: He/PRP ho/VBD pehreyen/VBN saal/JJ ware/NN class/NN maen/NN
10	اهي ڪله راند جي ميدان ۾ هئا	They were in the playground yesterday	Uhe huoa rand maidan mean kalh	Uhe, huoa, rand, maidan, mean, kalh	Tagged Text: Uhe/NNP huoa/NN rand/NN maidan/NN mean/NN kalh/NN

A. Analysis of the Parts-of-Speech Tagging

The output from the word labelling task of Romanized Sindhi text performed using the online SindhiNLP Python tool [9] and Sindhi rule-based model was analyzed in which 13 different POS categories were identified. The detailed statistics of the word labelling task are shown in Table IV.

TABLE IV. DETAIL STATISTICS FOR WORD LABELLING OF ROMANIZED SINDHI TEXT

Description	***Total Number of Words***	***Total number of POS Tagged Words***	***Word Labelling of Romanized Sindhi Text***	
Romanized Sindhi Text (100 sentences were used)	624	624	POS	No. of Words
			NNP	110
			NN	372
			PRP	11
			JJ	13
			RB	11
			WP	4
			VBD	54
			VBZ	0
			VBN	4
			VBP	30
			VB	4
			WDT	0
			DT	11
			Total	624

From the results produced by the SindhiNLP POS tagger, 624 Sindhi words was successfully tagged in which 482 are noun and pronouns, 92 verbs and 50 determinants were found as illustrated in Fig. 4.

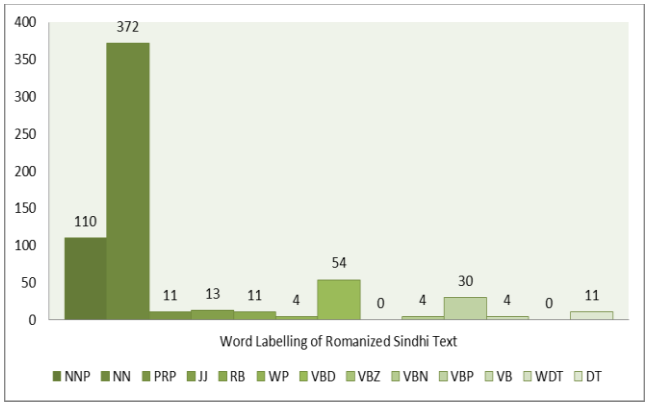


Fig. 4. Word Labelling of Romanized Sindhi Text.

V. CONCLUSION

In research study of Word by Word Labelling of Romanized Sindhi Text the conclusion is based on the following outcomes.

- A hybrid approach was used that combines online and manual approaches and a rule-based algorithm was designed and applied to the word labelling tasks.
- From the results 13 different POS categories were identified and 654 words of Romanized Sindhi Text were tested by using the SindhiNLP Python tool and all words were tagged successfully.
- From the results 482 noun/pronouns were found while the remaining 172 words were found to be adjectives, adverbs, verbs and determiners.
- For future work, Romanized Sindhi text from different domains will be used in the word labelling tasks and results will be compared using different machine learning techniques and tools.

REFERENCES

[1] Iyengar, Arvind. "A diachronic analysis of Sindhi multiscriptality." Journal of Historical Sociolinguistics 7, no. 2 (2021): 207-241.

[2] J. Lalwani, “History of Sindhi Language”, Voice of Sindhistaan, Vol. 4, no. 4, (2005). http://www.sindhishaan.com/article/language/lang_04_04.html

[3] History of Sindh - Govt. of Sindh [Retrieved on June 27, 2022]. https://www.sindh.gov.pk/history

[4] Nair, Jayashree, Riyaz Ahammed, and Anakha Shaji. "A Study on Transliteration Techniques and Conventional Transliteration Schemes for Indian Languages." In Sustainable Communication Networks and Application, pp. 103-117. Springer, Singapore, 2022.

[5] Ali, Wazir, Rajesh Kumar, Yong Dai, Jay Kumar, and Saifullah Tumrani. "Neural Joint Model for Part-of-Speech Tagging and Entity Extraction." In 2021 13th International Conference on Machine Learning and Computing, pp. 239-245. 2021.

[6] Saeed, Hafiz Hassaan, Muhammad Haseeb Ashraf, Faisal Kamiran, Asim Karim, and Toon Calders. "Roman Urdu toxic comment classification." Language Resources and Evaluation 55, no. 4 (2021): 971-996.

[7] AL MANSOORI, M. O. U. Z. A. "Exploring Sentiment Analysis using Different Machine Learning Algorithms on Dialectal Arabic." PhD diss., The British University in Dubai (BUiD), 2021.

[8] Arora, Gaurav. "iNLTK: Natural language toolkit for indic languages." arXiv preprint arXiv:2009.12534 (2020).

[9] Online Python tool http://text-processing.com/demo/

[10] Li, Hongwei, Hongyan Mao, and Jingzi Wang. "Part-of-Speech Tagging with Rule-Based Data Preprocessing and Transformer." Electronics 11, no. 1 (2021): 56.

[11] Sodhar, Irum Naz, Akhtar Hussain Jalbani, Muhammad Ibrahim Channa, and Dil Nawaz Hakro. "Parts of speech tagging of Romanized Sindhi text by applying rule based model." IJCSNS 19, no. 11 (2019): 91.

[12] Sodhar, Irum Naz, Akhtar Hussain Jalbani, Abdul Hafeez Buller, Muhammad Ibrahim Channa, and Dil Nawaz Hakro. "Sentiment analysis of Romanized Sindhi text." Journal of Intelligent & Fuzzy Systems 38, no. 5 (2020): 5877-5883.

[13] Sodhar, Irum Naz, Akhtar Hussain Jalbani, and Muhammad Ibrahim Channa. "Identification of issues and challenges in romanized Sindhi text." International Journal of Advanced Computer Science and Applications 10, no. 9 (2019).

[14] Abbasi, Muhammad Hassan, and Sajida Zaki. "LANGUAGE SHIFT: JOURNEY OF THIRD GENERATION SINDHI AND GUJARATI

SPEAKERS IN KARACHI." Bahria University Journal of Humanities & Social Sciences 2, no. 1 (2019): 19-19.

[15] Shackle, C. "Sindhi language." Encyclopedia Britannica, July 9, 2018. https://www.britannica.com/topic/Sindhi-language.

[16] Zeroual, Imad, Abdelhak Lakhouaja, and Rachid Belahbib. "Towards a standard Part of Speech tagset for the Arabic language." Journal of King Saud University-Computer and Information Sciences 29, no. 2 (2017): 171-178.

[17] Sodhar, Irum Naz, Akhtar Hussain Jalbani, Muhammad Ibrahim Channa, and Dil Nawaz Hakro. "Romanized Sindhi rules for text communication." Mehran University Research Journal Of Engineering & Technology 40, no. 2 (2021): 298-304.

[18] Afini, Umriya, Catur Supriyanto, and Raden Arief Nugroho. "The Development of Indonesian POS Tagging System for Computer-aided Independent Language Learning." International Journal of Emerging Technologies in Learning 12, no. 11 (2017).

[19] Ekbal, Asif, S. Mondal, and Sivaji Bandyopadhyay. "POS Tagging using HMM and Rule-based Chunking." The Proceedings of SPSAL 8, no. 1 (2007): 25-28.

[20] Devi, S. Anjali, and S. Sivakumar. "A Hybrid Ensemble Word Embedding based Classification Model for Multi-document Summarization Process on Large Multi-domain Document Sets." International Journal of Advanced Computer Science and Applications 12.9 (2021).

[21] Btoush, Mohammad Hjouj, Abdulsalam Alarabeyyat, and Isa Olab. "Rule based approach for Arabic part of speech tagging and name entity recognition." International Journal of Advanced Computer Science and Applications 7.6 (2016).

[22] Khan, Sadiq Nawaz, et al. "Urdu word segmentation using machine learning approaches." International Journal of Advanced Computer Science and Applications 9.6 (2018).

Forest Fires Detection using Deep Transfer Learning

Mimoun YANDOUZI[1]
Lab. LSI, ENSAO
Mohammed First University
Oujda, Morocco

Mounir GRARI[2]
Lab. MATSI, ESTO
Mohammed First University
Oujda, Morocco

Idriss IDRISSI[3]
Lab. MATSI, ESTO
Mohammed First University
Oujda, Morocco

Mohammed BOUKABOUS[4]
Lab. MATSI, ESTO
Mohammed First University
Oujda, Morocco

Omar MOUSSAOUI[5]
Lab. MATSI, ESTO
Mohammed First University
Oujda, Morocco

Mostafa AZIZI[6]
Lab. MATSI, ESTO
Mohammed First University
Oujda, Morocco

Kamal GHOUMID[7]
Lab. LSI, ENSAO
Mohammed First University
Oujda, Morocco

Aissa KERKOUR ELMIAD[8]
Lab. LARI, FSO
Mohammed First University
Oujda, Morocco

***Abstract*—Forests are vital ecosystems composed of various plant and animal species that have evolved over years to coexist. Such ecosystems are often threatened by wildfires that can start either naturally, as a result of lightning strikes, or unintentionally caused by humans. In general, human-caused fires are more severe and expensive to fight because they are frequently located in inaccessible areas. Wildfires can spread quickly and become extremely dangerous, causing damage to homes and facilities, as well as killing people and animals. Early discovery of wildfires is vital to protect lives, property, and resources. Reinforced imaging technologies can play a key role to detect wildfires earlier. By applying deep learning (DL) over a dataset of images (collected using drones, planes, and satellites), we target to automate the forest fire detection. In this paper, we focus on building a DL model specifically to detect wildfires using transfer learning techniques from the best pretrained DL computer vision architectures available nowadays, such as VGG16, VGG19, Inceptionv3, ResNet50, ResNet50V2, InceptionResNetV2, Xception, Dense-Net, MobileNet, MobileNetV2, and NASNetMobile. Our proposed approach attained a detection rate of more than 99.9% over multiple metrics, proving that it could be used in real-world forest fire detection applications.**

***Keywords*—Forest fires; wildfires; deep learning; transfer learning; computer vision; convolutional neural networks (CNN)**

I. Introduction

Forests are one of the most important natural resources on our planet. They support a diverse range of plants and animals' lives, play an important role in climate regulation, and provide numerous economic and social benefits. However, forests are vulnerable to damage and destruction, and wildfires are one of their most serious threats [1]. A wildfire can start accidentally (with a spark from a campfire), or it can be deliberately set by someone who intends to cause harm. Wildfires can quickly spread, destroying everything in their path. They can also cause significant environmental damage, such as the death of trees and other plants, the removal of soil, and the release of harmful emissions into the atmosphere [2].

In recent years, wildfires have become increasingly severe. Wildfires raged through Algeria, Tunisia, and Morocco in mid-July 2021, as well as Italy and Greece in the Mediterranean. The fires, which are believed to have been started by arsonists, burned through thousands of acres of land, killing dozens of people and injuring many more [3], [4]. The fires were especially devastating to the local economies, causing widespread agricultural damage as well as the destruction of businesses and homes. Furthermore, the tourism industry suffered as many people canceled their trips to the affected areas. Despite the efforts of firefighters and volunteers, the fires kept burning for weeks, leaving a trail of destruction in their wake.

We can do a lot to reduce the risk of wildfires, such as properly managing forests and using fire-resistant materials when building houses and other structures. In addition, we need to address the root cause of these fires.

The detection of wildfires at a preliminary phase is critical for protecting people, property, and resources. Imaging technology has the potential to aid in the early detection of wildfires. High-altitude drones, aircraft, and satellites can detect wildfire heat signatures by top shooting the fire area [5].

In this paper, we aim to build the most accurate DL model for forest fires using transfer learning out of the most achieving and well-known computer vision architectures pre-trained models available today, such as VGG, Inceptionv3, ResNet50, InceptionResNetV2, Xception, Dense-Net, MobileNet, and NASNetMobile.

The rest of the paper is organized as follows. The second section provides a focus on the used techniques, while the third section deals with the related works. Then the fourth section

presents our proposed method. Before concluding, the fifth section examines and discusses our study's findings.

II. Background

A. Computer Vision (CV)

CV is the process of extracting useful information from digital images. This data could be used for tasks such as object recognition, scene description, and motion tracking [6]. There are two kinds of computer vision algorithms: low-level and high-level. Low-level algorithms work on individual pixels; whereas, high-level ones work on more abstract features such as edges and corners. Low-level algorithms are typically faster and more accurate, but they are also more complex and require more processing power. High-level algorithms are less accurate, but they are also faster and easier to implement. On the other hand, deep learning has shown great promise in this area and has been used to achieve impressive results in tasks such as object recognition and scene understanding [7]. Deep learning-based computer vision can recognize objects in images, recognize faces, and read text. It can also be used for automatic image tagging and classification [8].

B. Convolutional Neural Networks (CNN)

CNNs are a type of deep learning (DL) network, which means they are made up of multiple layers of neurons organized in a hierarchical structure. A deep learning network's goal is to learn representations of input data by gradually extracting more and more information from it. Because of their ability to learn features of the input data, CNNs are particularly well-suited for computer vision tasks [9]. This is accomplished through a process known as feature extraction, which involves identifying the important features in the data and representing them in a way that the network can learn from. In contrast, traditional machine learning algorithms require the programmer to explicitly specify which features the algorithm should use [10].

CNNs have the ability to learn features that are specific to the task at hand, which is one of their advantages [11], [12]. It is made up of several layers, each with its own function. The first layer is the input layer, which receives input data in the form of a numerical value matrix and feeds it into a series of convolutional and pooling layers. This combination of convolutional and pooling layers is known as a kernel. The output layer is the final layer in a CNN; it produces the network's results (see Fig. 1).

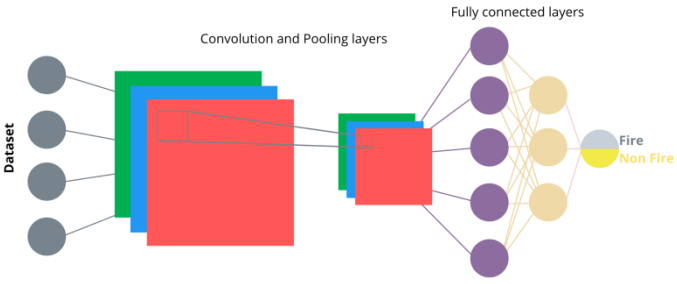


Fig. 1. CNN Layers Architecture.

We list below the most successful deep learning CNN architectures for image recognition:

- VGG is a deep learning architecture (created by the Visual Geometry Group) that was also created for the ILSVRC contest (ImageNet Large Scale Visual Recognition Challenge). The VGG16 is made up of 16 layers, including five convolutional layers, four dense layers, and a final fully connected layer, while the VGG19 model has 19 layers [13].
- Inception is a deep learning model that performed well in the ImageNet competition. It is made up of a deep convolutional network with a large number of layers (more than 20) [14].
- ResNet (Residual neural network) is a deep neural network that performed well in the ILSVRC. It is made up of a deep convolutional network with a large number of layers (more than 100), one of ResNet's major purposes is a so-called “identity shortcut connection” that skips one or more layers [15].
- InceptionResNetV2 is a variant of the original Inception-v2 model, designed to improve its performance on the ImageNet dataset. The model is composed of an Inception module followed by a ResNet module [16].
- Xception (Xtreme Inception) is a deep learning model, based on a CNN with a large number of layers (more than 150) [17].
- DenseNet is a CNN model with a large number of layers (more than 500), designed to increase the number of connections between neurons. This helps to improve the overall accuracy of the network.[18].
- MobileNet is a deep learning framework that enables developers to create sophisticated neural networks for mobile devices. It is designed to be efficient and lightweight, making it suitable for running on a wide range of mobile devices. MobileNet50 version has a depth of 50 layers and can be used for both classification and detection tasks [19].
- NASNet (Neural Architecture Search Network) is a CNN model trained on the ImageNet dataset [20]. It automates network architecture engineering by searching for the best algorithm to achieve the best performance on a certain task, while automatically configuring the number of layers, the number and type of neurons in each layer, and the architecture of the network. The NASNetMobile version is suitable for mobile devices [21].

III. Related Work

Dutta S. et al [22] proposed a hybrid architecture of separable convolution neural networks and digital image processing employing thresholding and segmentation for reliably detecting small-scale forest burning, which generally heralds the beginning of more terrible catastrophes. Performance examination of the test data on the suggested design provided outstanding results in terms of high sensitivity (98.10 %) and specificity (87.09 %).

Aslan S. et al [23] proposed a smoke detection approach based on Deep Convolutional Generative Adversarial Neural Networks (DC-GANs). In order to ensure a robust representation of sequences with and without smoke, the training framework includes regular training of a DCGAN with real pictures and noise vectors, as well as training the discriminator separately using smoke images without the generator. With a TNR of 99.45% and a TPR of 86.23%, the suggested approach is able to identify smoke pictures in real time with minimal false positives.

Wang Y. et al [24] proposed a forest fire image identification system based on traditional image processing methods and convolutional neural networks, and an adaptive pooling methodology was established to identify fire automatically. Using this technique, the features of the fire flame may be segmented and learned in advance. It has been shown in experiments that the adaptive pooling convolutional neural network approach has greater performance and a higher recognition rate, with an accuracy as high as 90.7%.

Chen Y. et al [25] proposed a UAV-based forest fire detection approach based on a convolutional neural network method in order to identify a probable fire in its early stages. Experimentation with generated flames in an indoor testbed proves that the suggested fire detection system works.

We will make a comparative study between a wide range of deep learning models, practically all of those that have demonstrated their effectiveness in computer vision, in order to propose the most accurate model possible for forest fire detection

IV. Proposed Method

Our proposed solution is to detect wildfires before they spread out of control using drones and deep learning algorithms. Drones are used to fly over forests and identify hot spots that could spark a wildfire. Once a hot spot has been identified, the deep learning algorithm can be used to determine whether it is a wildfire, and notify the authorities via a cloud server (Fig. 2).

This scheme has several advantages over traditional methods. First, drones can fly over large areas much faster than ground crews. Second, the deep learning algorithm can identify wildfires much more accurately than human observers can. Third, the use of drones and deep learning algorithms can help to protect firefighters and their assistants and keep them safe from danger by alerting them earlier.

Our main contribution in this paper is to build the most accurate DL model specifically for detecting wildfires in forests from the best resulting DL computer vision architectures available at the time, by leveraging previously known knowledge from pre-trained models. These models are already trained to know certain categories, and we narrow their knowledge to focus only on two categories (Fire or Non-Fire).

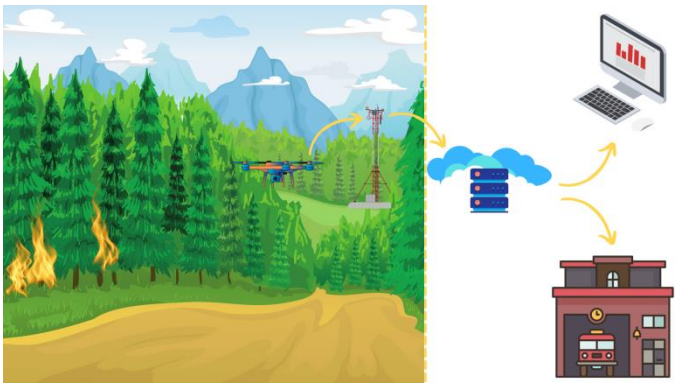


Fig. 2. Components and Functions of a UAV-Based Forest Fire Detection Platform.

A. Dataset

A common challenge in deep learning is obtaining datasets that are sufficiently large and diverse in nature for the task at hand [26], [27]. The dataset used for training our models is comprised of a large number of images captured of wildfires in different locations around the world, as well as images of forest landscapes with no fire. It was constructed by mixing and merging multiple smaller datasets from search engines and Kaggle [28], resulting in 4661 images in our new dataset; 2525 images with the label "no fire" and 2136 images with the label "fire", after cleaning some corrupted images.

In addition, we performed data augmentation on the dataset, allowing us to significantly increase the size of our training dataset and, as a result, the quality of the trained models [29]. With data augmentation, we added new data to the dataset that is similar to the original data, but with some slight modifications, which can improve the performance of neural networks learning from data and improve their accuracy [30]. We used different data augmentation techniques [31], such as:

1) Random rotation: this can help to reduce overfitting by creating new images that are rotated versions of existing images. This also gives the model a chance to learn how to recognize objects from different angles.

2) Horizontal and vertical mirroring: they can also help to reduce overfitting by providing the model with new images that are mirror images of existing images. This can also help the model learn to recognize objects that may be upside down or rotated in different orientations.

3) Gaussian blur: it can help to improve the robustness of the model by making the images less detailed and more forgiving of small changes. This can help the model to generalize better to new data.

4) Pixel level augmentation: it can help to improve the model's ability to learn from small changes in the input data. This can be useful for learning from data that may be noisy or have low resolution.

B. Building the Models

A model can be trained in a variety of ways. In this section, we will stare at transferring pre-trained models to a new task. Transfer learning models are typically constructed by first training them on a large dataset, such as the ImageNet dataset. This model is then used as the "base model" for another model trained on a smaller dataset (see Fig. 3). In our case, the smaller dataset is often a more specialized dataset (images of fires and forests). The smaller model is then tuned to better fit the dataset on which it is being used. This process is frequently repeated, with the final model trained on a dataset even smaller than the original. This process of model training is commonly referred to as fine-tuning [32]. We used this same strategy for each of the state-of-the-art models, importing the pre-trained DL model class [33], while ensuring that we can add our own custom input and output layers according to our data. While leveraging the previously learned weights during the initial training on the old data, a massive amount of time and space is saved while minimizing the model's complexity. (see Fig. 3).

Afterward, we inserted a fully connected and output layer (new classifier) after the pre-trained model was imported so that new real learning could take place; the fully connected layer is a flatten layer and a dense layer with 512 neurons [34]. The sigmoid activation function is used in the output layer with only one output neuron matching the binary label in our data (Fire or Not). Finally, we train our models on 80% of the augmented new dataset (Training set) and validate the obtained results with the remaining 20% (Validation set).

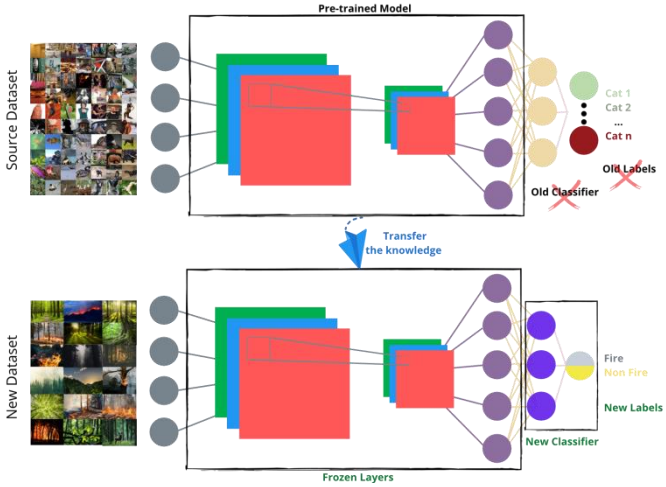


Fig. 3. Transfer Learning Technique.

V. Results and Discussions

A. Hardware and Software Characteristics

In order to get our results, we used TensorFlow on an HPC system with the following hardware specifications:

- 2x Intel Gold 6148 (2.4 GHz/20 cores) CPUs
- 2x NVIDIA Tesla V100 graphics cards, each having 32GB of RAM

TensorFlow v2.7.0 was used in our experiments, it is an open-source data analysis and machine learning software library. It was first developed by engineers and researchers at the Google Brain team in 2015. TensorFlow provides a wide range of capabilities for data analysis and machine learning, including numerical computing, linear algebra, graph processing, and deep learning [35].

B. Evaluation Metrics

It is necessary to have a proper evaluation metric in place in order to find the best model during the training phase [36]. When evaluating deep learning models, certain metrics must be used, such as Accuracy, Precision, Recall, and Loss. In order to calculate these metrics, four different parameters are used [37]:

- True Positive (TP): is the total of successfully categorized positive class records.
- True Negative (TN): is the total of successfully categorized negative class records.
- False Positive (FP): is the total of incorrectly categorized positive class records.
- False Negative (FN): is the total of incorrectly categorized negative class records.

1) Accuracy: Accuracy is the percentage of correctly classified items. It is the most basic and common evaluation metric for classification tasks. The accuracy is simply the ratio of correctly predicted labels out of all predicted labels [34]:

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

2) Loss: It is a measure of how far off the algorithm is from the desired output. The lower the loss, the better the algorithm is performing [38]. The cross-entropy loss is a commonly used loss metric for classification problems. It is calculated by this formula:

$$Loss = -\sum_i y_i * \log(p_i) \quad (2)$$

Where y_i is the output of the true label and p_i is its predicted probability. The cross-entropy loss is used to assess the performance of a classifier by penalizing incorrect predictions.

The higher the cross-entropy loss, the more incorrect predictions the classifier is making.

3) Precision: Precision is a metric estimating how well a model predicts true positives. True positives are those instances that are correctly identified as positive by the model [39].

A model with high precision will correctly identify the most positive examples, while a model with low precision will misclassify many positive examples as negative. The Precision metric is given by:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

4) Recall: The recall is the ratio of correctly predicted positive instances to all positive instances (see formula (4)). It is also known as the true positive rate or sensitivity. A model with high recall is capable of detecting the most positive

instances [39]. A model with low recall is not informative as it classifies most positive instances as negative.

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

In general, Recall should be used alongside other metrics such as Precision and Accuracy to get a complete picture of a model's performance.

5) The number of parameters: The number of parameters in deep learning refers to the number of variables that are used to define the structure of the neural network. These variables can be the weights and biases of the network, the size of the network, or the type of activation function used. It reflects the learning capacity of the model; A deep learning model with a large number of parameters has the capacity to learn more complex patterns than a model with fewer parameters [40]. The number of parameters also determines the amount of memory required to store the model; A model with a large number of parameters requires more memory than a model with fewer parameters [41]. The number of parameters in the models will vary from the original pre-trained models owing to the change in the fully connected layers (the convolution base was unmodified as it was frozen).

C. Evaluating the Results

In Table I and Figures 4-7, we present the obtained results on multiple metrics; the accuracy, loss, precision, and recall, along with the number of parameters for each model (this number differs from the original pre-trained models, due to our new classifier). To maximize effectiveness, we trained all of the models over one hundred epochs. These obtained results show that the ResNet50, VGG16, and VGG19 algorithms have higher accuracies, lower losses, and higher recalls and precisions, achieving a near-perfect score. Meanwhile, MobileNet and DenseNet came in second place with more than 97% in three metrics (accuracy, precision, and recall), but with a loss of around 5 to 6%; On the other hand, MobileNetV2 achieves close results, more than 96% in the three metrics and a loss of more than 9%. Then, Xception, which received more than 94% in the three metrics, and a high Loss averaging 14 to 15%. ResNet50V2 obtained mediocre results; even though its first version (ResNet50) got good results, around 84-85% in the three metrics, but with high losses up to 34% (higher errors are related to high loss, which means that the model does not do a good job). NASNetMobile and InceptionV3 performed similarly to ResNet50V2 in all metrics. On the other side, the mixed-model InceptionResNetV2 performed the worst in the accuracy, precision, and loss metrics, but reached the best score in the recall metric (100%). This shows that the model has a low false-negative rate (down to zero), but with a high false-positive rate due to the low precision results. At the end of this discussion, the best models retained are ResNet50, VGG16, and VGG19. Then, we compared their number of parameters. They have the respective numbers: ResNet50 (24.6~ million), VGG16 (~14.9 million), and VGG19 (~20.2 million). if we are looking for a model with the best learning capacity, ResNet50 is the accurate candidate; On the other hand, if we are targeting a lightweight model to deploy on a limited resource and battery-connected devices such as a drone or an IoT thing [42], VGG16 is the suitable one among the three. It has60% fewer parameters in comparison with the ResNet50. With fewer performances, DenseNet is the best lightweight model (after VGG16) with only 7.5 million parameters. Also, if we prioritize model size, MobileNet will be the best choice with only 3,7m parameters and an accuracy close to 98%, MobileNetV2 is the lightest model in this case study with only 2,9m with a modest accuracy of more than 96% just a little behind it first version MobileNet.

For the other models, ResNet50V2 has about the same number of parameters as ResNet50, while Xception and InceptionV3 have respectively ~24.6m and ~21.9m, but produced modest results. Despite its high number of parameters (55.1m), InceptionResNetV2 is the poorest model in our case study, indicating that deeper networks or more neurons do not always produce the best results.

TABLE I. ACHIEVED RESULTS FOR THE IMPLEMENTED MODELS

Deep Learning Algorithm	Number of parameters	Accuracy	Loss	Precision	Recall
VGG16	14.977.857	99.81%	0.49%	99.77%	99.89%
VGG19	20.287.553	99.78%	0.48%	99.83%	99.78%
InceptionV3	22.852.385	83.23%	37.48%	83.47%	87.21%
ResNet50	24.637.313	99.94%	0.19%	99.94%	99.94%
ResNet50V2	24.614.401	84.89%	34.89%	85.20%	87.68%
Inception ResNetV2	55.124.193	55.19%	68.78%	55.19%	100%
Xception	21.911.081	94.09%	15.54%	94.58%	94.84%
DenseNet	7.562.817	97.50%	6.88%	97.84%	97.62%
MobileNet	3.754.177	97.87%	5.24%	98.02%	98.13%
MobileNetV2	2.914.369	96.28%	9.53%	96.74%	96.47%
NASNetMobile	4.811.413	85.09%	33.14%	84.13%	89.98%

NASNet Mobile is the lightweight model in our case study (~4.8m) it is an edge devices model however its performance is insufficient for our purposes.

Fig. 8 shows predicted image samples that demonstrate that our system can almost perfectly distinguish between fire and normal forest state regardless of all the features and variety of objects (people. snow. different types of trees. etc.). Fig. 9 shows the incorrectly predicted images using the VGG16 and ResNet50 models (these images are collected from the Web and not seen by the model neither in training nor in validation) which can easily explain why the model wrongfully predicted the bad labels. The most likely explanation for the negative results is that they can really deceive the human eye; in the first image the sun and its radiation in the clouds and the lake can be easily misinterpreted as fire because we see the same

features and patterns of flames. In the second and third images our system incorrectly misidentifies fog as fire.

No system is perfect, but these findings show that deep learning can be extremely accurate in detecting wildfires, with a success rate of more than 99.9%. Thanks to its ability to identify the unique signatures emitted by wildfires, this can provide an early warning of a wildfire, allowing fire crews to be dispatched to the scene before it spreads too far. This solution could be a valuable tool for fire departments and other emergency responders in identifying and responding to wildfires.

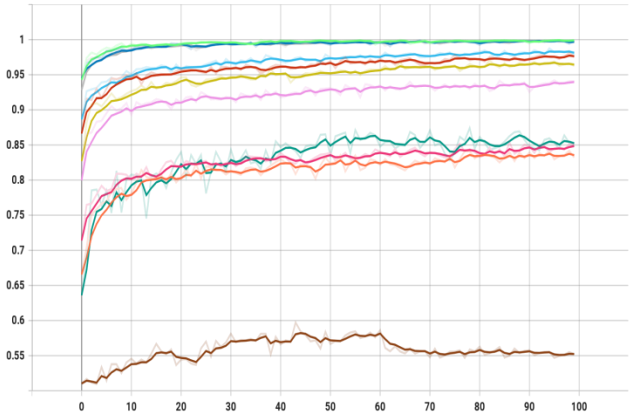


Fig. 4. Achieved Accuracy (over 100 Epochs) for the Implemented Models.

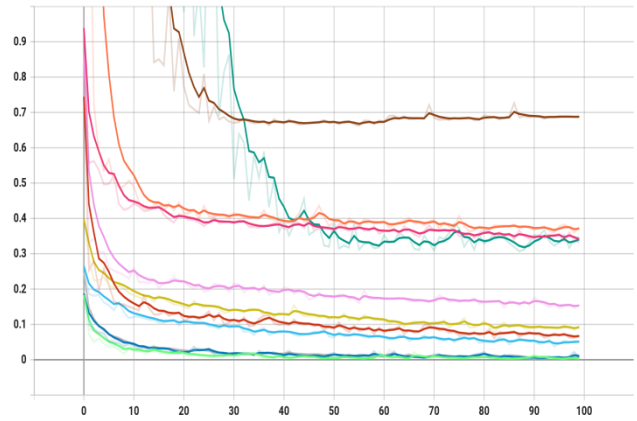


Fig. 5. Achieved Loss (over 100 Epochs) for the Implemented Models.

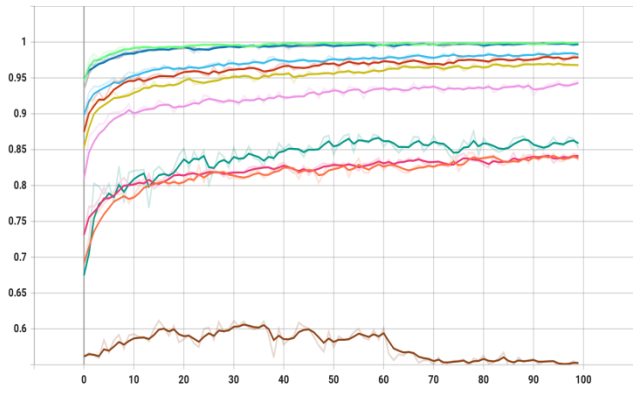


Fig. 6. Achieved Precision (over 100 Epochs) for the Implemented Models.

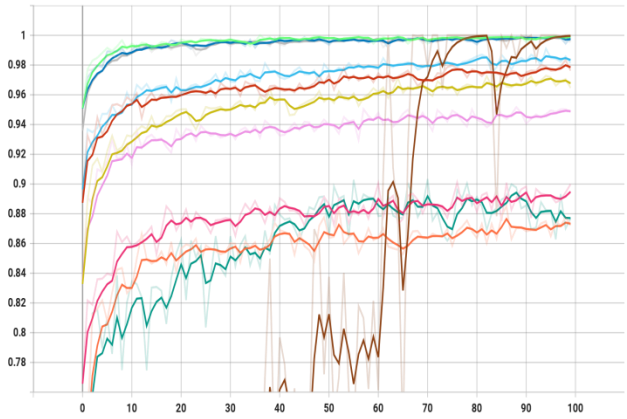


Fig. 7. Achieved Recall (over 100 Epochs) for the Implemented Models.

Fig. 8. Examples of Predicted Wildfires.

Fig. 9. Examples of Wrongly Predicted Wildfires.

In the future, we will deal with the incorrect negative cases, in which the system can be confused between flames and the sun, fog and clouds, and smoke. Furthermore, we will try to

implement these models as the feature extractor backbone in other DL algorithms such as the R-CNN (Region-Based Convolutional Neural Networks) family [43], SSD (Single Shot Detector) [44], or applying YOLO (You Only Look Once) [45], [46] in order to detect not only fires but also its precise coordinates.

VI. Conclusion

Deep learning has revolutionized computer vision by enabling computers to learn from data to recognize patterns and classify objects with high accuracy. This has led to the development of powerful computer vision algorithms and applications that can detect and identify objects in photos and videos with a high degree of accuracy. Our proposed approach in this paper involves building a deep learning model specifically for detecting wildfires in forests using the transfer learning technique. Our discussion based on the obtained results has given us VGG16 and ResNet50 as relevant models for our issue; they are able to achieve higher scores in accuracy, recall and precision of more than 99.9% and a loss down to 0.19% for ResNet50 and down to 0.48% for VGG16. Fire departments and other emergency responders may benefit from these techniques to better identify and control wildfires before they spread too far. Through future works we will try to improve and develop these models by using object detection approaches such as the R-CNN family, SSD and YOLO to identify fires based on their precise location coordinates.

Acknowledgment

This work is supported by the Mohammed First University under the PARA1 Program (Low-cost and real-time Forest Fire Detection System based on Wireless Sensor Networks - SDF-RCSF). The used computational resources of HPC-MARWAN are provided by the National Center for Scientific and Technical Research (CNRST). Rabat. Morocco.

References

[1] M. Grari, I. Idrissi, M. Boukabous, O. Moussaoui, M. Azizi, and M. Moussaoui, “Early wildfire detection using machine learning model deployed in the fog/edge layers of IoT,” Indones. J. Electr. Eng. Comput. Sci., vol. 27, no. 2, 2022.

[2] M. Yandouzi et al., “Review on forest fires detection and prediction using deep learning and drones,” J. Theor. Appl. Inf. Technol., vol. 100, no. 12, pp. 4565–4576, 2022.

[3] Africanews, “Wildfires bring devastation to Algeria, Tunisia | Africanews.” https://www.africanews.com/2021/08/12/wildfires-bring-devastation-to-algeria-tunisia/ (accessed Jan. 26, 2022).

[4] Africanews, “Forest fires rage in northern Morocco | Africanews.” https://www.africanews.com/2021/08/16/forest-fires-rage-in-northern-morocco/ (accessed Jan. 26, 2022).

[5] M. Grari et al., “Using IoT and ML for Forest Fire Detection, Monitoring, and Prediction: a Literature Review,” J. Theor. Appl. Inf. Technol., vol. 100, 2022.

[6] J. Peters, “Foundations of Computer Vision - Computational Geometry, Visual Image Structures and Object Shape Detection,” 2017, pp. 1–443. Accessed: Jan. 26, 2022. [Online]. Available: https://books.google.com/books/about/Foundations_of_Computer_Vision.html?hl=fr&id=CtxmDgAAQBAJ

[7] A. Kherraki, M. Maqbool, and R. El Ouazzani, “Traffic Scene Semantic Segmentation by Using Several Deep Convolutional Neural Networks,” 2021 3rd IEEE Middle East North Africa Commun. Conf., pp. 1–6, Dec. 2021, doi: 10.1109/MENACOMM50742.2021.9678270.

[8] A. Kherraki and R. El Ouazzani, “Deep convolutional neural networks architecture for an efficient emergency vehicle classification in real-time traffic monitoring,” IAES Int. J. Artif. Intell., vol. 11, no. 1, pp. 110–120, Mar. 2022.

[9] M. Boukabous and M. Azizi, “Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers,” Indones. J. Electr. Eng. Comput. Sci., vol. 25, no. 2, Feb. 2022, doi: 10.11591/IJEECS.V25.I2.PP.

[10] I. Idrissi, M. Azizi, and O. Moussaoui, “IoT security with Deep Learning-based Intrusion Detection Systems: A systematic literature review,” in 4th International Conference on Intelligent Computing in Data Sciences, ICDS 2020, Nov. 2020, pp. 1–10. doi: 10.1109/ICDS50568.2020.9268713.

[11] M. Berrahal and M. Azizi, “Augmented Binary Multi-Labeled CNN for Practical Facial Attribute Classification,” Indones. J. Electr. Eng. Comput. Sci., vol. 23, no. 2, pp. 973–979, Aug. 2021.

[12] I. Idrissi, M. Boukabous, M. Azizi, O. Moussaoui, and H. El Fadili, “Toward a deep learning-based intrusion detection system for IoT against botnet attacks,” IAES Int. J. Artif. Intell., vol. 10, no. 1, pp. 110–120, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp110-120.

[13] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., Sep. 2014, Accessed: Jan. 26, 2022. [Online]. Available: https://arxiv.org/abs/1409.1556v6

[14] C. Szegedy et al., “Going Deeper with Convolutions,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June-2015, pp. 1–9, Sep. 2014, doi: 10.1109/CVPR.2015.7298594.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.

[16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” 31st AAAI Conf. Artif. Intell. AAAI 2017, pp. 4278–4284, Feb. 2016, Accessed: Feb. 12, 2022. [Online]. Available: https://arxiv.org/abs/1602.07261v2

[17] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 1800–1807, Oct. 2016, doi: 10.1109/CVPR.2017.195.

[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 2261–2269, Aug. 2016, doi: 10.1109/CVPR.2017.243.

[19] A. G. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” arXiv:1704.04861, Apr. 2017, Accessed: Jul. 14, 2022. [Online]. Available: http://arxiv.org/abs/1704.04861

[20] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning Transferable Architectures for Scalable Image Recognition,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 8697–8710, Jul. 2017, doi: 10.1109/CVPR.2018.00907.

[21] B. Zoph and Q. V. Le, “Neural Architecture Search with Reinforcement Learning,” 5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc., Nov. 2016, Accessed: Feb. 12, 2022. [Online]. Available: https://arxiv.org/abs/1611.01578v2

[22] S. Dutta and S. Ghosh, “Forest Fire Detection Using Combined Architecture of Separable Convolution and Image Processing,” 2021 1st Int. Conf. Artif. Intell. Data Anal. CAIDA 2021, pp. 36–41, Apr. 2021, doi: 10.1109/CAIDA51941.2021.9425170.

[23] S. Aslan, U. Gudukbay, B. U. Toreyin, and A. Enis Cetin, “Early Wildfire Smoke Detection Based on Motion-based Geometric Image Transformation and Deep Convolutional Generative Adversarial Networks,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2019-May, pp. 8315–8319, May 2019, doi: 10.1109/ICASSP.2019.8683629.

[24] Y. Wang, L. Dang, and J. Ren, “Forest fire image recognition based on convolutional neural network:,” https://doi.org/10.1177/1748302619887689, vol. 13, Nov. 2019, doi: 10.1177/1748302619887689.

[25] Y. Chen, Y. Zhang, J. Xin, Y. Yi, D. Liu, and H. Liu, "A UAV-based Forest Fire Detection Algorithm Using Convolutional Neural Network," Chinese Control Conf. CCC, vol. 2018-July, pp. 10305–10310, Oct. 2018, doi: 10.23919/CHICC.2018.8484035.

[26] I. Idrissi, M. Azizi, and O. Moussaoui, "An unsupervised generative adversarial network based-host intrusion detection system for internet of things devices," Indones. J. Electr. Eng. Comput. Sci., vol. 25, no. 2, pp. 1140–1150, Feb. 2022, doi: 10.11591/IJEECS.V25.I2.PP1140-1150.

[27] M. Boukabous and M. Azizi, "Review of Learning-Based Techniques of Sentiment Analysis for Security Purposes," in Innovations in Smart Cities Applications Volume 4, Springer, Cham, 2021, pp. 96–109. doi: doi.org/10.1007/978-3-030-66840-2_8.

[28] "Forest Fire Images | Kaggle." https://www.kaggle.com/mohnishsaiprasad/forest-fire-images (accessed Jan. 27, 2022).

[29] M. Berrahal and M. Azizi, "Optimal text-to-image synthesis model for generating portrait images using generative adversarial network techniques," Indones. J. Electr. Eng. Comput. Sci., vol. 25, no. 2, Feb. 2022, doi: 10.11591/IJEECS.V25.I2.PP.

[30] M. Berrahal and M. Azizi, "Review of DL-Based Generation Techniques of Augmented Images using Portraits Specification," in 4th International Conference on Intelligent Computing in Data Sciences, ICDS 2020, Nov. 2020, pp. 1–8. doi: 10.1109/ICDS50568.2020.9268710.

[31] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," 2018 Int. Interdiscip. PhD Work. IIPhDW 2018, pp. 117–122, Jun. 2018, doi: 10.1109/IIPHDW.2018.8388338.

[32] M. Boukabous and M. Azizi, "A comparative study of deep learning based language representation learning models," Indones. J. Electr. Eng. Comput. Sci., vol. 22, no. 2, pp. 1032–1040, 2021, doi: 10.11591/ijeecs.v22.i2.pp1032-1040.

[33] "Keras Applications." https://keras.io/api/applications/ (accessed Jan. 30, 2022).

[34] I. Idrissi, M. Azizi, and O. Moussaoui, "Accelerating the update of a DL-based IDS for IoT using deep transfer learning," Indones. J. Electr. Eng. Comput. Sci., vol. 23, no. 2, pp. 1059–1067, Aug. 2021, doi: 10.11591/ijeecs.v23.i2.pp1059-1067.

[35] "API Documentation| TensorFlow Core v2.7.0." https://www.tensorflow.org/api_docs (accessed Jan. 27, 2022).

[36] M. Hossin and Sulaiman, "A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS," IJDKP) Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, 2020, doi: 10.5121/ijdkp.2015.5201.

[37] "Metrics to Evaluate your Machine Learning Algorithm | by Aditya Mishra | Towards Data Science." https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 (accessed Sep. 13, 2020).

[38] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in Advances in Neural Information Processing Systems, 2018, vol. 2018-Decem, pp. 8778–8788.

[39] F. Cakir, K. He, X. Xia, B. Kulis, and S. Sclaroff, "Deep Metric Learning to Rank." pp. 1861–1870, 2019.

[40] M. Geiger et al., "Scaling description of generalization with number of parameters in deep learning," J. Stat. Mech. Theory Exp., vol. 2020, no. 2, p. 023401, Feb. 2020, doi: 10.1088/1742-5468/AB633C.

[41] "The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems." https://proceedings.neurips.cc/paper/1991/hash/d64a340bcb633f536d56e51874281454-Abstract.html (accessed Mar. 17, 2022).

[42] I. Idrissi, M. Mostafa Azizi, and O. Moussaoui, "A Lightweight Optimized Deep Learning-based Host-Intrusion Detection System Deployed on the Edge for IoT," Int. J. Comput. Digit. Syst., vol. 11, no. 1, pp. 209–216, 2022, doi: 10.12785/ijcds/110117.

[43] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 1, pp. 142–158, Jan. 2016, doi: 10.1109/TPAMI.2015.2437384.

[44] W. Liu et al., "SSD: Single Shot MultiBox Detector," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9905 LNCS, pp. 21–37, Dec. 2015, doi: 10.1007/978-3-319-46448-0_2.

[45] "YOLO: Real-Time Object Detection." https://pjreddie.com/darknet/yolo/ (accessed Jul. 26, 2020).

[46] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv:1804.02767v1, 2018, Accessed: Jul. 26, 2020. [Online]. Available: https://pjreddie.com/yolo/.

An Enhancement Technique to Diagnose Colon and Lung Cancer by using Double CLAHE and Deep Learning

Nora yahia Ibrahim, Amira Samy Talaat
Computers and Systems Department
Electronics Research Institute, Cairo, Egypt

***Abstract*—The most common and deadly cancers are lung and colon cancers. More than a quarter of all cancer cases are caused by them. Early detection of the disease, on the other hand, greatly raises the probability of survival. Image enhancement by Double CLAHE stages and modified neural networks are made to improve classification accuracy and use Deep Learning (DL) algorithms to automate cancer detection. A new Artificial Intelligent classification system is presented in this research to recognize five kinds of colon and lung tissues, three malignant and two benign, with three classes for lung cancer and two classes for colon cancer, based on histological images. The results of the study imply that the suggested system can accurately identify tissues of cancer up to 99.5%. The use of this model will aid medical professionals in the development of an automatic and reliable system for detecting different kinds of colon and lung tumors.**

***Keywords*—Artificial intelligent system; machine learning; cancer detection; image classification; deep learning**

I. Introduction

Cancer is one of the top causes of death worldwide, according to the Organization of World Health. Autonomous growth, genetic instability, and substantial metastatic potential are acquired by cancer cells. Colon and lung are the most affected organs, with the largest number of deaths. Colon cancer is the major cause of 9.2% of cancer mortality worldwide, while lung cancer is the major cause of 18.4% of all cancer mortality [1, 2]. The combined frequency of colon and lung cancer is estimated to be around 17%. Although this is improbable, cancer cell spread across these two organs is highly common in the absence of early diagnosis [3].

Only effective treatment and early detection can now minimise cancer deaths [4]. The faster a patient is diagnosed, the more effective the treatment and the better the patient's chances of survival and healing.

To search for cancer cells and rule out other probable diseases, many tests are performed, including sputum cytology, imaging sets (CT scan, x-ray), and biopsy (tissue sampling). The examination of microscopic histopathology slides by trained pathologists while performing the biopsy is important in determining the diagnosis [5, 6] and identifying tumour forms and subsets [7]. This study uses just histopathology images to diagnose colon and lung cancers automatically.

Health specialists frequently employ histopathological images for analysis, and they are crucial in determining the survival chances of patients. Usually, health specialists had to go through a lengthy process to diagnose cancer by reviewing histopathological images. However, with the technological tools accessible now, this process may be completed with less time and effort [3]. Artificial intelligence systems have recently gained popularity for their capability to analyze data quickly and give conclusions.

II. Literature Review

In biomedical applications, machine learning techniques are used to predict and classify various types of signals and images. Machines can now deal with large-scale data such as anatomical multidimensional videos and images because of deep learning (DL) methods. Deep learning is a machine learning field that builds algorithms to produce an artificial neural network built on the human brain's structure and function [8]. The majority of previous research used DL to categorise lung and colon cancer images simultaneously. Some writers concentrated on colon cancer detection, while others concentrated on lung cancer detection.

A deep learning-based algorithm is used by Masud [9] to classify colon and lung histological images. They used two types of domain modifications to obtain four image categorization feature sets. They joined the properties of the two categories to arrive at the final categorization result. They were 96.33% accurate. By employing a shallow neural network design, Mangal [10] was able to classify colon and lung cancers based on histological images. In classifying lung and colon malignancies, they reached an accuracy of 97% and 96%, respectively.

Hatuwal [11] proposed a deep learning method based on CNN. In the method, they present samples of only lung tissues from the dataset. This approach could only identify two malignant and one benign tissue in the lung, and no information on colon cancer categorization was given. Their suggested lung tissue categorization model attained an accuracy of 97.20%, a recall of 97.33%, and a precision of 97.33%. Sarwinda [12] suggested a classifier of KNN with characteristics retrieved for colon tissues by a DenseNet-121 pretrained network. Their approach mines the information for colon tissues and distinguishes between benign and malignant tissues of the colon. For colon categorization, their

model achieved 98.53% accuracy and 98.63% recall. Their model, however, was unable to collect lung tissues and provided no information about lung classification. According to Kumar [13], DenseNet-121 extracts more significant characteristics than other CNN pre-trained networks. This is because of the use of small links to improve the accuracy and efficiency of the network. Wang [14] built a Python library based on deep learning to detect cancer image categories. In their proposed strategy, they combined the CNN model and the SVM algorithm. The SVM model's overall accuracy was 94%.

Chehade [15] identifies colon and lung cancer subtypes, and the model of XGBoost offers the best classification rate in terms of recall, accuracy, and precision. XGBoost had a 99% accuracy and a 98.8% F1 score.

Hlavcheva [16] employed convolutional neural networks to analyze medical images using deep learning techniques. The dataset was used to compare the accuracy of several CNN designs in classification. The accuracy of 94.6% was achieved using neural network theory and statistical mathematical methodologies.

The study's primary goal is to develop a medical analytical intelligent support system for colon and lung imaging and, using machine learning, develop an automated method for properly classifying the subtypes of lung and colon cancer from histopathological images so we can achieve high levels of accuracy.

The following is a summary of the contributions of this paper:

- We proposed a novel colon and lung Image classification technique by applying the Image enhancement technique combination of DWT (discrete wavelet transform) and Double-CLAHE (Double Contrast Limited Adaptive Histogram Equalization) in the Preprocessing phase of the image.
- CLAHE is applied twice, first for the low frequency decomposed part of the Image DWT component and after the inverse DWT of the reconstructed image, which makes image details more enhanced.
- We proposed a new hybrid combination of an enhanced image from DWT with Double-CLAHE, EfficientNetB7 Deep learning technique, and adding Modified Neural Network method to fully discover the multi-class deep-broad characteristics of the colon and lung Image dataset.
- The proposed method demonstrates outstanding improvement in the performance for the training and testing datasets and gives a very high classification accuracy of 99.5%.

The following is how the article is organized. Section III discusses the datasets on the colon and lungs. Section IV Methodology with implementation details and results evaluation. Section V of Experimental results, comparisons, and conclusion

III. Dataset on Colon and Lung

The proposed technique is tested using the LC25000 dataset [17], a new colon and lung cancer histopathology image dataset that was published in 2020. This collection, which was put up by Andrew A. Borkowski and his colleagues, has 25000 colour images of five lung and colon tissues of different types [18], namely Benign Colonic Tissue, Benign Lung Tissue, Colon Adenocarcinoma, Lung Squamous Cell Carcinoma, and Lung Adenocarcinoma. Table I shows the details of the dataset as well as the allocated class names.

TABLE I. The Details of the LC25000 Database

Cancer Type	***Name of Category***	***Number of Images***
Colonic _Benign _Tissue	Col_Be	5000
Lung _Benign _Tissue	Lun_Be	5000
Colon _Adenocarcinoma	Col_Ad	5000
Lung _ Carcinoma _Squamous _Cell	Lun_Sc	5000
Lung _Adenocarcinoma	Lun_Ad	5000
Total	5	25000

Adenocarcinoma is the most frequent type of colon cancer, accounting for more than 95% of all cases. When a form of polyp (tissue growth) called an adenoma grows in the large intestine, it becomes an adenocarcinoma and progresses to cancer. Lung adenocarcinoma makes up around 40% of all lung tumors, and it affects more females than males. This form of cancer generally starts in glandular cells and spreads to the lungs' alveoli. All tumours that grow in the colon and lungs are not malignant and do not travel to other regions of the body.

These tumours are classified as benign, and they aren't usually fatal. They must, however, be removed surgically and biopsied to determine if malignancy is present. Lastly, lung carcinoma squamous cell is a type of small cell tumour that arises in the airways or bronchi of the lungs. It is the second most frequent kind of lung cancer, accounting for roughly 30% of all cases. Only 500 images of the colon and 750 images of the lung are included in the original LC25000 dataset. They enlarged the dataset to 25,000 images by using augmentation strategies to flip and rotate the original images under various situations (each class has 5000 images).

The original images were 1024 x 768 pixels in size. However, to make them square, they were resized to 768 x 768 pixels before using the augmentation methods. Sample histopathology images from the LC25000 dataset from these five classes are shown in Fig. 1.

IV. Methodology

This section describes the suggested deep learning-based classification method for colon and lung cancer diagnosis. The Convolutional Neural Network is a method for distinguishing cancers from other cells or tissues that has been shown to be effective [19].

The EfficientNet-B7 network was fine-tuned in this paper to classify colon and lung tumours. Histopathology images are

shown in Fig. 2. The proposed method structure consists of three main stages: image pre-processing with enhancement, EfficientNetB7, and the Modified Neural Network (MNN) stages, as shown in Fig. 2.

EfficientNetB7 takes the resized images from the first stage Fig. 3 to train these images for solving the classification problem. EfficientNetB7 is also used to extract the feature maps of colon and lung cancer histopathology images. However, the MNN takes the extracted feature from the second stage as input and a class label as output.

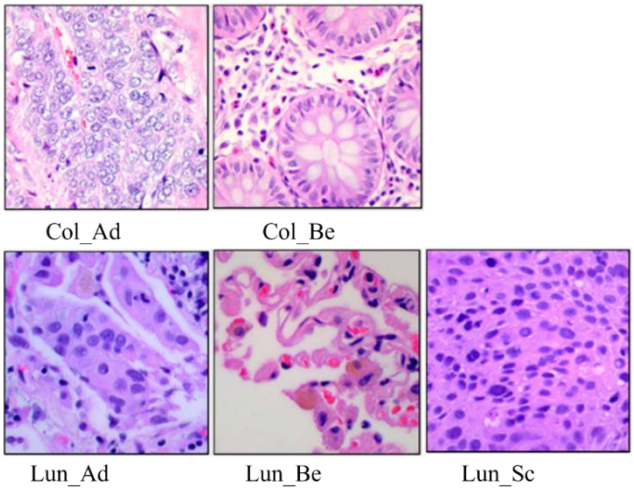


Fig. 1. LC25000 Dataset Sample Images.

The details of the proposed method with MBConv (Mobile Inverted Bottleneck Convolution), the resolution, number of channels, number of levels of each feature map, and the details of the MNN stage are also shown in Fig. 2. In the following subsection, each stage will be described in detail.

A. *Stage of Image Pre-Processing using DWT and Double-CLAHE*

The pre-processing image stage is required before the feature extraction procedure to prepare and clarify the images with labels for training the model.

Blurriness, poor border recognition, artifacts, and overlapping problems in histopathology images were caused by uneven staining of the slide because of human error.

As shown in Fig. 3, The Double-CLAHE approach (Double Contrast Limited Adaptive Histogram Equalization) is intended to eliminate these types of imperfections or uneven staining. The CLAHE method improves image contrast by increasing poor boundary edges in each pixel of an image through restricted amplification [20], as well as improving local contrast in an image. As a result, it's ideal for enhancing the features of histopathological images. This paper proposes a new image enhancing method that combines CLAHE and DWT (Discrete Wavelet Transform). Preprocessing of images in Fig. 2 was done using the DWT and CLAHE approaches in Fig. 3.

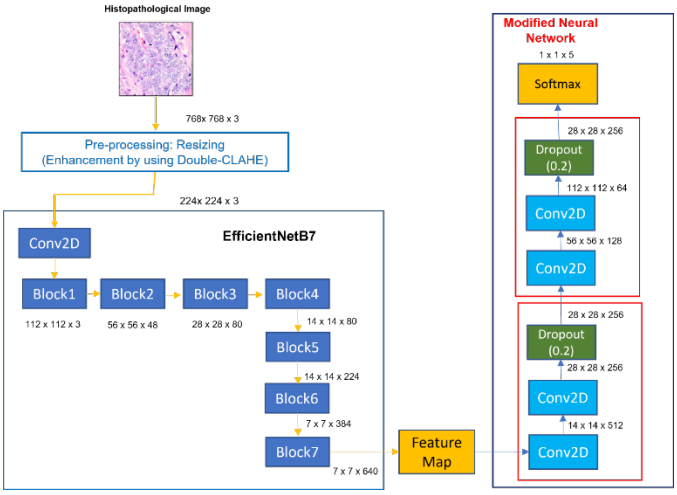


Fig. 2. The Proposed Framework for Image Classification.

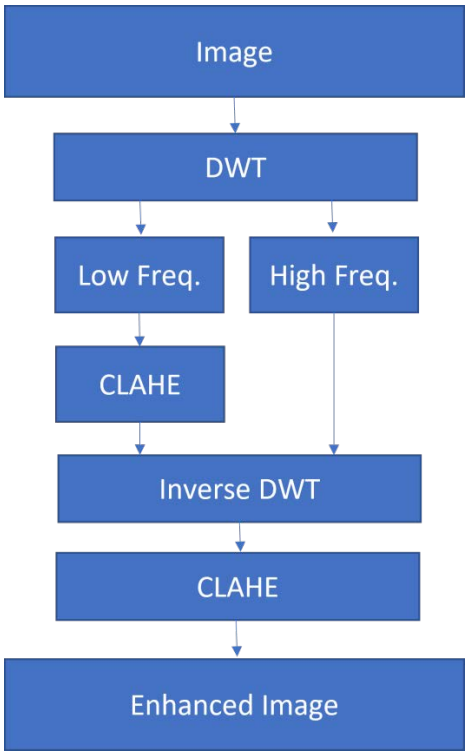


Fig. 3. The Steps of Enhancement by DWT and Double-CLAHE.

The new technique consists of four major steps: The original image is first decomposed into high-frequency and low-frequency components by DWT. The low-frequency values are then boosted by CLAHE while the high-frequency values are left untouched to limit noise amplification. This is because the high-frequency component refers to detailed information and comprises the majority of the original image's noise. Third, reconstruct the image using the inverse DWT of the new coefficients. Finally, after obtaining the reconstructed image, CLAHE is required to enhance the image in order to make the details more abundant.

The colon and lung cancer histopathology images are resized from 768x768 to 224x224 in RGB format to train the suggested model with the dataset.

B. Stage of EfficientNetB7

One of the most powerful CNN structures is EfficientNet. It employs a compound scaling strategy to increase network depth, width, and resolution, resulting in good capacity in a variety of benchmark datasets while using fewer computational resources than other models [21].

EfficientNets come in eight different models, from EfficientNet-B0 to EfficientNet-B7. The simplest model, EfficientNet-B0, is designed automatically by the Neural Architecture Search. Using the compound scaling method, the EfficientNet family is created by scaling up EfficientNetB0. Scaling the network increases model performance by balancing all architecture image resolution, depth, width, and compound coefficients.

Excitation optimization and squeeze in mobile inverted bottleneck convolution (MBConv) [22] is the core of the EfficientNet architecture. Fig. 4 depicts the MBConv concept.

The number of MBConv blocks in the EfficientNet network family varies. The depth, width, resolution, and model size keep increasing as EfficientNetB0 through EfficientNetB7 improve, as well as the accuracy [21]. Efficient-NetB7 exceeds previous CNNs on ImageNet in terms of accuracy, and it is furthermore 6.1x faster and 8.4x smaller than the best available CNN [21]. MBConv is the fundamental building block of the network. The filter size identifies each MBConvX block. It corresponds to X=1 and X=6 which represent the standard ReLU and ReLU6 activation functions, respectively. Fig. 5 depicts the characteristics of the seven blocks' architecture.

Flattening the extracted feature-maps yields a single vector of features once the features are extracted from the dataset images. A Modified Neural Network stage takes this vector as its input.

C. The Modified Neural Network Stage

It is the last stage of the proposed method of the classification process. As clarified in Fig. 2, the Modified Neural Network stage contains four convolution layers, two dropout layers, and a softmax layer. The softmax layer (output layer) of the proposed method is customized with the number of our classes.

The aim of this phase is to add variety to the extracted knowledge and assist it to have a better understanding of the samples, allowing them to be categorized more accurately.

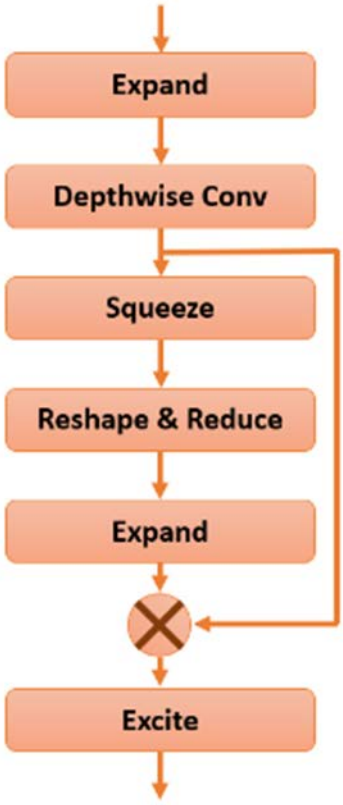


Fig. 4. EfficientNet Basic Building Block (MBConv).

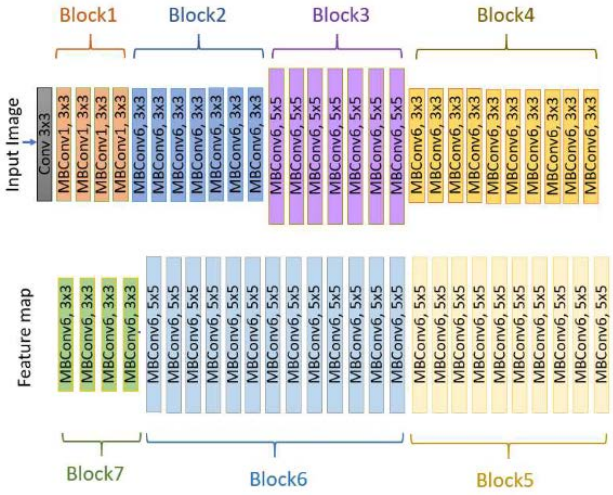


Fig. 5. The Architecture of Block1 to Block 7 of EfficientNetB7.

V. Experimental Results Using

A. The Implemented Details

The obtained outcomes of the implemented experiment are mentioned in this section. The input dataset is split into 85:15, with 85% of the images (randomly chosen) for training and

15% for validating. Because our dataset is balanced (every class contains the same number of images), the system will be less subject to bias while making decisions.

TABLE II. THE IMPLEMENTED DETAILS OF THE PROPOSED CNN MODEL FOR CLASSIFICATION TASK

Variable	***Value***
Image dimensions	224 x 224
Initial channels	3
Dropout	20%
Batch Size	64
Epochs	22
Convolutional layer activation	Relu
Learning rate	0.001
activation of Dense layer	Softmax
Compiler-optimizer	Adam
Compiler-loss	Categorical-cross-entropy

The proposed model was developed using Tensorflow 2.0. As shown in Table II, the system is trained on an image with a size of 768x768 pixels by scaling it to 224x224 pixels, using a batch size of 64 and 22 epochs. For initializing the training, the weights that have been pre-trained by EfficientNetB7 on ImageNet are used, and they are fine-tuned. For training, the ADAM optimizer with a learning rate of 0.001 and a categorical-cross-entropy loss function is utilised. The proposed framework's performance in a classification problem is measured using accuracy, average precision (AP), average recall (AR), and the F1 measure, which will be discussed in the next section.

B. Evaluation of Performance

Machine learning models are evaluated using a variety of criteria. The confusion matrix, as well as associated metric factors like precision, F1-score, accuracy, and recall, are utilized to measure in this paper.

The classifier's accuracy is a measure of its capability to correctly classify instances. It refers to the percentage of valid results or correctly identified samples among all samples.

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

The True Negative, True Positive, False Negative, and False Positive values are represented by TP, TN, FP, and FN, respectively. TP denotes true disease, i.e., the true value is positive, and it is classified positively, indicating that the patient has the disease and that the test is positive.

A false Negative (FN) shows that the patient has the disease while the test is negative, suggesting that the real value is positive but the classification is negative. A False positive (FP) denotes the presence of a disease when none exists, implying that the real value is negative when classed positively. A True Negative (TN) denotes that the patient is healthy and the test is negative, signifying that the true value is negative, and the test is negative.

Precision is denoted as the proportion of correctly identified samples (true positives) to positive samples identified.

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

Recall is the percent of positive samples of a specific class that are accurately identified. The proportion of real positive samples to total positive samples is used to compute it.

$$\text{Rec}all=\frac{TP}{TP+FN} \quad (3)$$

The F1-score is known as the harmonic average of accuracy and recall.

$$\text{F1-score}=\frac{2\times Precision\times Recall}{Precision+Recall}$$

C. Results

The outcome of the proposed framework for the classification of colon and lung cancer histological images is shown in this section. To determine their performance, the models were evaluated using test data.

Fig. 6 shows the proposed model's each epoch classification accuracy. The experiment has 22 epochs in all. The classification accuracy on the testing subset was 99.36% at the last epoch; however, the greatest results were at epochs 12, 13, and 14, all of which had a 99.47% accuracy. The training accuracy curve increased gradually and almost steadily towards the top, as shown in the figure.

At epoch number 18, the greatest training accuracy was 99.5%, which is quite similar to the accuracy of the previous epoch 99.36%. The curve of testing accuracy is similar to the training accuracy curve, with the outcome improving as the training progresses. At 20 epochs, the curve drops to 96.7%, indicating that the model is able to give a satisfactory classification result even if it is constructed with fewer epochs.

Fig. 7, on the other hand, shows the training and validation loss, which represents the percentage of data loss for each classification attempt.

As seen in Fig. 7, both the training and validation subsets' loss values decreased as the number of epochs grew.

The normalized (ROC) Receiver Operating Characteristic and confusion matrix curves of the testing subset classification at the 18th epoch are shown in Fig. 8. For the test data given labelled categories, the confusion matrix compares the images' true labels against their predicted labels. Only 3% (112 samples) of the testing images (3750 samples) were misclassified, as shown by the normalized confusion matrix.

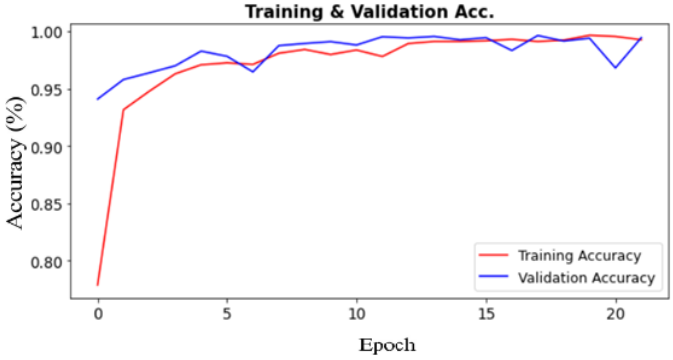


Fig. 6. The Visual Display of the Accuracy Rate of the Proposed Classification Model at Each Epoch.

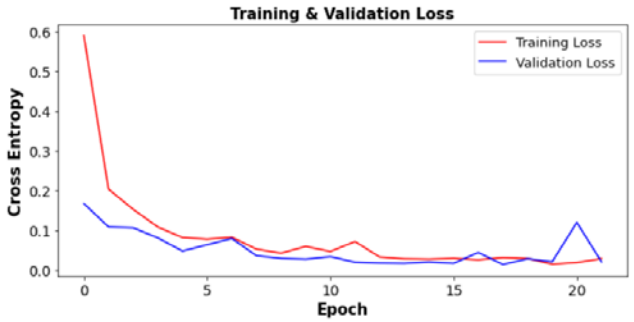


Fig. 7. The Proposed Classification System's Optimal Training and Validation Loss.

The best classification results are in the Col_Be and Lun_Be categories, while the other categories, Lun_Ad, Col_Ad, and Lun_Sc, have the same misclassification rate. These results can also be seen in the ROC curves.

Because the classifier was quite successful at separating the samples, the Lun_Be and Col_Be curves have reached the top-left corner. Overall, the suggested deep learning approach is highly precise in classifying these classes.

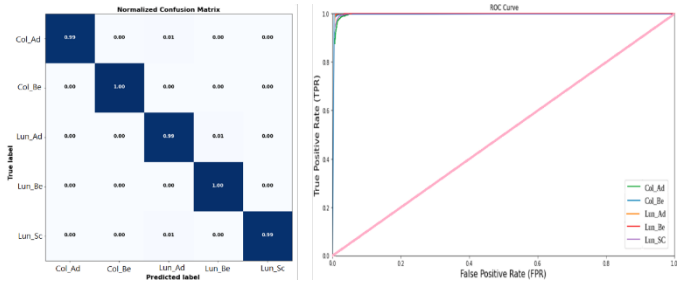


Fig. 8. Representations of Classification Results: (a) Normalized Confusion Matrix (b) Epoch 18th ROC Curve.

Table III displays the recall, F1-score, and precision of the proposed classification model on the test data for five classes of histological images. Table III shows that the average recall, precision, and F1-score for each of the five categories is more than 0.994. In addition, except for Lun_Sc, our classification approach attained the highest precision in all classes.

TABLE III. THE RECALL, PRECISION, AND F1-SCORE FOR HISTOLOGICAL IMAGES OF COLON AND LUNG CANCER

Categories	*Precision*	*Recall*	*F1-score*
Col_Ad	**1.00**	0.995	**1.00**
Col_Be	**1.00**	**1.00**	**1.00**
Lun_Ad	**1.00**	0.98	0.99
Lun_Be	**1.00**	**1.00**	**1.00**
Lun_Sc	0.98	**1.00**	0.99
Average	0.996	0.994	0.996

D. Comparison

We compare our model to current models in the literature, which are given in the introduction, to evaluate the proposed method. Table IV compares the results of the lung and colon cancer subtype classification with other approaches using the same dataset. As shown in Table IV, our system outperforms existing cancer detection technologies in terms of maximal classification accuracy.

TABLE IV. COMPARISON OF THE RESULTS FOR THE SAME DATASET WITH OTHER METHODS

References	*ClassifierModel*	*(%) Accuracy*	*(%) Precision*	*(%) Recall*	*(%) F1-score*
Masud, et al. [9]	CNN	96.33	96.39	96.37	96.38
Mangal [10]	CNN for lung Cancer	97.89	-	-	-
Mangal [10]	CNN for colon Cancer	96.61	-	-	-
Hatuwal [11]	CNN for lung Cancer	97.20	97.33	97.33	0.96
Sarwinda [12]	DenseNet-121-KNN for colon Cancer	98.53	-	98.63	-
Kumar [13]	DenseNet-121-DF	98.60	98.63	98.60	-
Wang [14]	CNN & SVM	94			90
Chehade [15]	XGBoost	99	98.6	99	98.8
Hlavcheva [16]	CNN-D	94.6	-	-	-
Proposed Method	The proposed classifier	**99.5**	**99.6**	**99.4**	**99.6**

E. Conclusion and Future Work

A deep learning technique is presented in this paper to classify images and will help us detect colon and lung cancer more precisely in the future. For this study, we utilised a histopathology image dataset that is freely accessible on Kaggle [17]. The model training accuracy achieved is 99.5% for the colon and lung dataset.

This paper proposed a method for colon and lung cancer classification problems. The method has two main sections: the enhancement of images by DWT and Double-CLAHE stages; and the modified neural network to enhance classification accuracy.

The result shows that: accuracy 99.5%, precision 99.6%, recall 99.4%, and F1-score 99.6%, which proves that the presented method is effective for solving colon and lung cancer classification problems. It also outperforms the previous approaches in terms of performance.

The optimized method that provided improved accuracy must be used in upcoming models. In the future, combining YOLO, 3D-CNN, and a variety of other approaches that are applied to various image datasets will allow us to construct more powerful and effective models in the future.

REFERENCES

[1] Bray, F., et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 2018. 68(6): p. 394-424.

[2] Bermúdez, A., et al., Her2-Positive and Microsatellite Instability Status in Gastric Cancer—Clinicopathological Implications. Diagnostics, 2021. 11(6): p. 944.

[3] Toğaçar, M., Disease type detection in lung and colon cancer images using the complement approach of inefficient sets. Computers in Biology and Medicine, 2021. 137: p. 104827.

[4] Sánchez-Peralta, L.F., et al., Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. Artificial intelligence in medicine, 2020. 108: p. 101923.

[5] Travis, W.D., et al., International association for the study of lung cancer/american thoracic society/european respiratory society

international multidisciplinary classification of lung adenocarcinoma. Journal of thoracic oncology, 2011. 6(2): p. 244-285.

[6] Abou Taleb, A.S.T. and A.F. Atiya, A new approach for leukemia identification based on cepstral analysis and wavelet transform. Int J Adv Comput Sci Appl, 2017. 8(7): p. 226-232.

[7] Yu, K., et al., Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun. 2016; 7: 12474. Epub 2016/08/17., https://doi.org/10.1038/ncomms12474 PMID: 27527408.

[8] Schmidhuber, J., Deep learning in neural networks: An overview. Neural networks, 2015. 61: p. 85-117.

[9] Masud, M., et al., A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. Sensors, 2021. 21(3): p. 748.

[10] Mangal, S., A. Chaurasia, and A. Khajanchi, Convolution Neural Networks for diagnosing colon and lung cancer histopathological images. arXiv preprint arXiv:2009.03878, 2020.

[11] Hatuwal, B.K. and H.C. Thapa, Lung cancer detection using convolutional neural network on histopathological images. Int. J. Comput. Trends Technol, 2020. 68: p. 21-24.

[12] Sarwinda, D., et al. Analysis of Deep Feature Extraction for Colorectal Cancer Detection. in 2020 4th International Conference on Informatics and Computational Sciences (ICICoS). 2020. IEEE.

[13] Kumar, N., et al., An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images. Biomedical Signal Processing and Control, 2022. 75: p. 103596.

[14] Wang, Y., et al., OCTID: a one-class learning-based Python package for tumor image detection. Bioinformatics, 2021. 37(21): p. 3986-3988.

[15] Chehade, A.H., et al., Lung and Colon Cancer Classification Using Medical Imaging: A Feature Engineering Approach. 2022.

[16] Hlavcheva, D., et al. Comparison of CNNs for Lung Biopsy Images Classification. in 2021 IEEE 3rd Ukraine Conference on Electrical and Computer Engineering (UKRCON). 2021. IEEE.

[17] Images, L.a.C.C.H., Kaggle. Available online: https://www.kaggle.com/andrewmvd/lung-and-colon-cancer-histopathological-images, 16 July 2020.

[18] Borkowski, A.A., et al., Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142, 2019.

[19] Kermany, D.S., et al., Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 2018. 172(5): p. 1122-1131. e9.

[20] Zuiderveld, K., Contrast limited adaptive histogram equalization. Graphics gems, 1994: p. 474-485.

[21] Tan, M. and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. in International conference on machine learning. 2019. PMLR.

[22] Howard, A., et al., Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. 2018.

Mobile Applications for the Implementation of Health Control against Covid-19 in Educational Centers, a Systematic Review of the Literature

Bryan Quispe-Lavalle[1], Fernando Sierra-Liñan[2], Michael Cabanillas-Carbonell[3]
Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú[1]
Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú[2]
Vicerrectorado de Investigación, Universidad Privada Norbert Wiener, Lima, Perú[3]

***Abstract*—A health crisis caused by the SARS-CoV-2 virus is still ongoing. That is why an important factor for the resumption of on-site classes is the creation of sanitary measures to help control Covid-19. The present research is a literature review, The PRISMA methodology is used and 265 articles are collected from various databases such as EBSCO Host, IEEE Xplore, SAGE, ScienceDirect, and Scopus. According to the inclusion and exclusion criteria, the most relevant articles aligned to the topic were identified, systematizing 119 articles. Showcasing digital technologies used in mobile applications that allow better control, tracking, and monitoring of the health status of students, teachers, and staff of educational centers, in addition to the parameters and quality attributes that must be taken into account for the effective sanitary control of the disease, finally, a development model is proposed.**

***Keywords*—Mobile application; sanitary control; systematic review; digital technologies**

I. Introduction

Covid-19 is caused by the severe acute respiratory syndrome called SARS-CoV-2. One of its most common symptoms is a lung infection or pneumonia [1]. Covid-19 is an epidemic that has spread rapidly throughout the world. That is why we must be alert to information on how to take care to prevent contagion [2]. The World Health Organization declared the coronavirus (Covid-19) a pandemic on March 11, 2020. Making all countries take preventive measures against the emerging Covid-19 virus [3].

Due to the distancing measures by Covid-19, education went from being face-to-face to virtual, therefore, schools, universities, and institutes closed their doors, and teachers and students had to adapt to the use of technological tools, making way for e-learning [4], [5].

Another very important area in which Covid-19 has had an impact has been in the area of life and mental health [6]. The impact was greater on students, as there were school closures, fear generation due to Covid-19, the interruption and change of modality in education, and the excessive use of digital devices. These factors have caused students to suffer from mental health problems such as stress, anxiety, depression, and sleep disorders during the quarantine period. With the vaccines, it was possible to reduce the mortality rate of the disease, and eventually return to face-to-face teaching in schools, so it is important to know the best digital technologies used for the control and monitoring of the virus.

The objective of this literature review is to analyze articles in order to have a better understanding of the problem, that is, to know which technologies, parameters, and quality attributes were used to have better sanitary control against Covid-19 in educational centers, as well as to identify the countries with more experiences in this field. Section II shows the methodology used in the search and selection of articles, Section III shows the results obtained through graphs and tables, Section IV the discussion in which the research questions posed are answered, and Section V shows the proposed model to be followed in future research, to finally write in Section VI the conclusions of the research.

II. Methodology

The methodology used consisted of three steps. First, the PRISMA methodology was used [7] (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Preferred reporting elements for systematic reviews and meta-analyses) which helped to find and identify the most appropriate articles for the present literature review. Second, bibliometric analysis was used to find the common terms that influence the disclosure of the implementation of health surveillance against Covid-19 in schools using digital technology. Finally, the most important factors and statistical methods used for the implementation of a sanitary control against Covid-19 in educational centers are extracted and related to the results of the bibliometric analysis.

Following the PRISMA methodology [7], this section is structured as follows: (1) Type of study, (2) Research questions, (3) Search strategy, and (4) Inclusion and exclusion criteria.

A. Type of Study

A systematic review of the literature will be used to prepare the article.

B. Research Questions

The proposed research questions are as follows:

RQ1. Which digital technologies allow better control, follow-up, and monitoring against Covid-19 of the health status of students, teachers, and staff in educational centers?

RQ2. What parameters should be taken into account to make effective sanitary control against Covid-19 in educational centers through the use of a mobile application?

RQ3. What quality attributes must it contain for the viability of the mobile application for the implementation of a health control against Covid-19 in educational centers?

RQ4. Which countries have the most research, in the last three years, related to health monitoring against Covid-19 in schools?

C. Search Strategies

To answer the research questions, a search for published articles was conducted in the main databases EBSCO Host, IEEE Xplore, SAGE, ScienceDirect, and Scopus. A total of 265 scientific articles were collected.

At the time of applying the search for our research, the following keywords were considered: "Covid-19" AND ("health control in schools" OR "in the schools" OR "using mobile application" OR "health control in schools app" OR "in the schools using mobile application" OR "prevention and control in schools"), "mobile application for" AND ("prevent Covid-19" OR "control in schools during Covid-19" OR "Covid-19 health control in schools" OR "prevent Covid-19" OR "the school to prevent Covid-19"), "app mobile for" AND ("health" OR "Covid-19" OR "Covid-19 in the schools"), "app mobile for Covid-19 in schools", "app control of Covid-19 in schools" y "control of Covid-19 in schools". The item collection process is shown in Fig. 1.

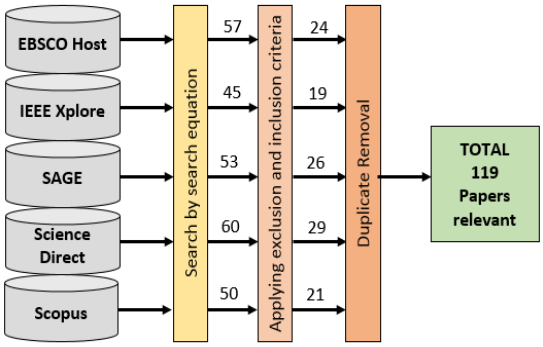


Fig. 1. Item Inclusion Chart.

D. Inclusion and Exclusion Criteria

For the systematic review study, the following inclusion and exclusion criteria were applied, as shown in Table I.

TABLE I. INCLUSION AND EXCLUSION CRITERIA

CRITERIA		
Inclusion	I01	Articles related to digital technologies for Covid-19 preventive control.
	I02	Articles published since the start of Covid-19 2019 – 2022.
	I03	Articles that consider at least one prevention parameter against Covid-19.
Exclusion	E01	Articles not related to digital technologies for preventive control against Covid-19.
	E02	Articles published before 2019.
	E03	Articles related to Covid-19 but do not make use of digital technologies.

III. RESULTS

A total of 265 articles found in the databases related to the research topic were analyzed, of which two duplicate articles were discarded or did not contribute similar topics. After reviewing the articles, 119 were selected, excluding 144 according to the exclusion criteria and which did not contribute to answering the research question. Obtaining 119 articles for the systematic review. Fig. 2 shows the selection process following the Prisma methodology.

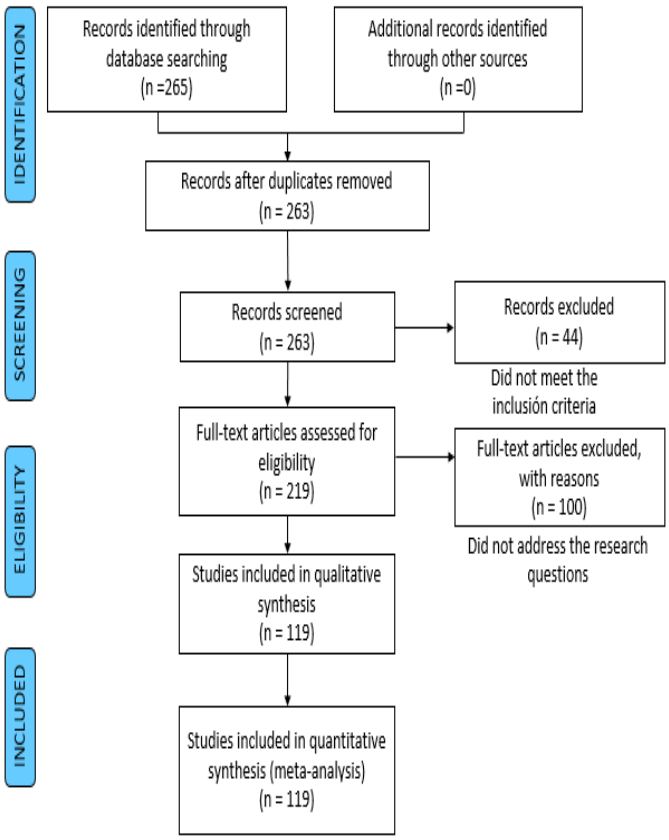


Fig. 2. PRISMA Diagram Methodology.

Fig. 3 shows the number of articles found by the database.

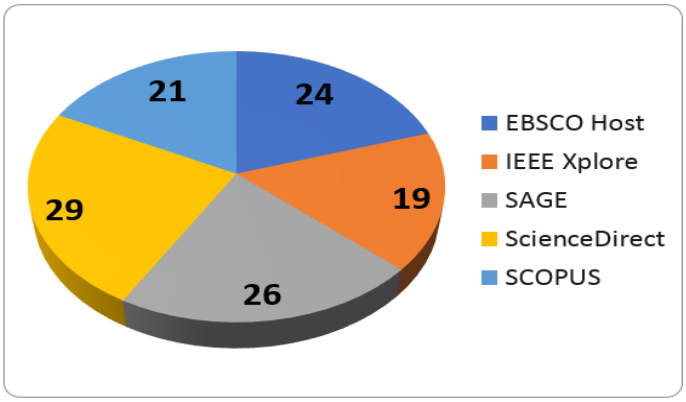


Fig. 3. Articles by Database.

Fig. 4 shows the number of articles published by year and database, selected in times of pandemic.

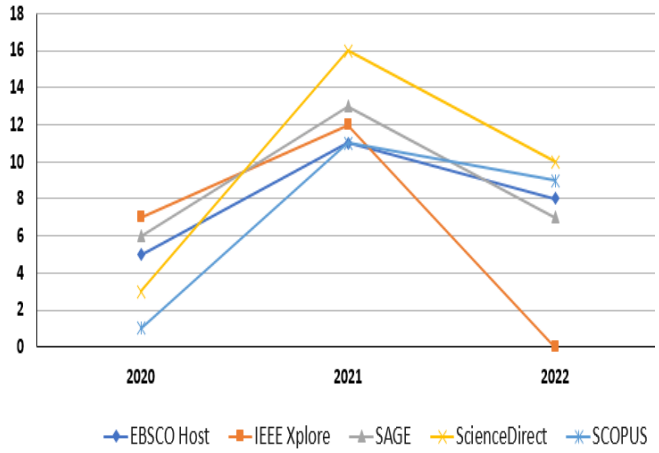


Fig. 4. Articles by Year and Database.

Fig. 5 shows the number of articles published by continent.

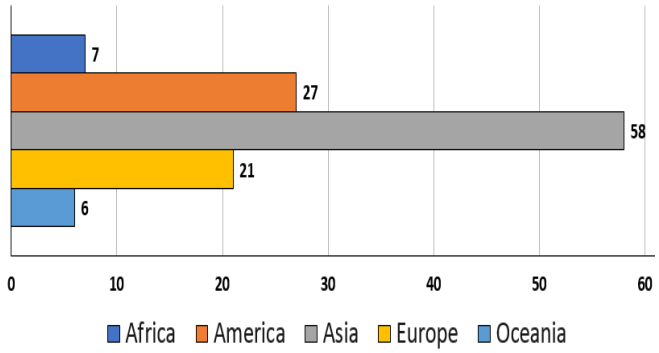


Fig. 5. Articles by Continent.

The number of articles published by country is shown in Fig. 6 on a scale from 1 to 17.

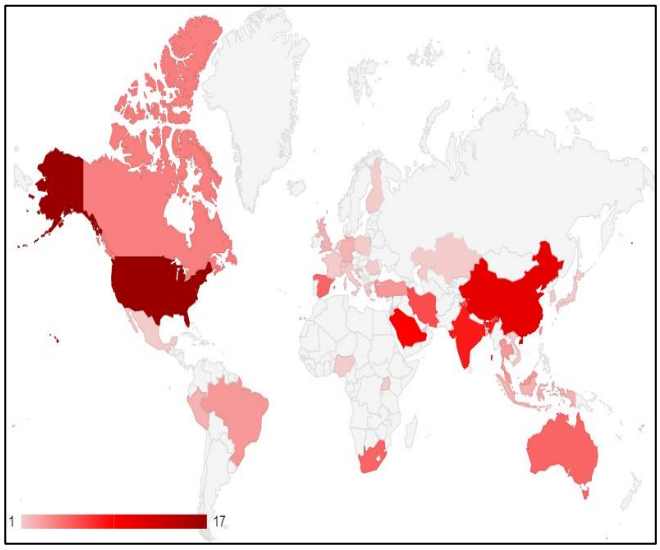


Fig. 6. Articles by Country.

Fig. 7 shows the network visualization based on a bibliometric analysis filtered by keywords, using the VOSviewer software.

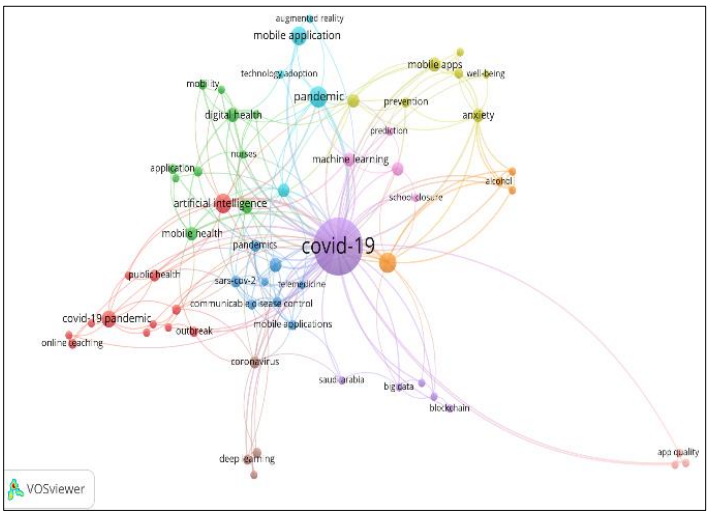


Fig. 7. Network Visualization of Bibliometric Analysis.

Fig. 8 shows the analysis of bibliometric data considering the year of publication of each article.

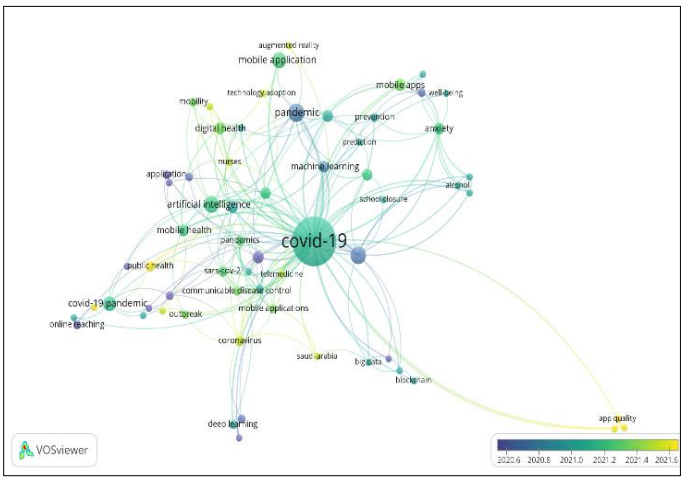


Fig. 8. Overlay Visualization of Bibliometric Analysis

It was in 1926 when Alfred Lotka introduced the term "bibliometrics" by analyzing the production patterns of different authors, concluding with the presentation of the first criteria for bibliometrics [8]. Bibliometrics is part of scientific research, as time goes by scientists become interested in this field and even academic institutions use it in their research work. It is a very effective technique to retrieve, evaluate and analyze, in a statistical way, quantifiable data in the academic literature, merits of a particular thematic area, or a particular publication containing indicators to obtain a better evolution of the research direction. Bibliometric analysis is expected to contribute to filling gaps in the research field, provide new perspectives for future research and promote collaboration [9].

VOSviewer is a software tool that allows us to construct and visualize bibliometric networks (including individual publications, authors, and scientific journals); it is constructed from co-authorship relationships, co-citation, bibliographic coupling, citation networks, and co-occurrence of important terms extracted from a body of scientific literature [10].

VOSViewer was used to obtain networks based on co-occurrences of important terms, from which visualization maps were created as shown in Fig. 7, Fig. 8, and Fig. 9.

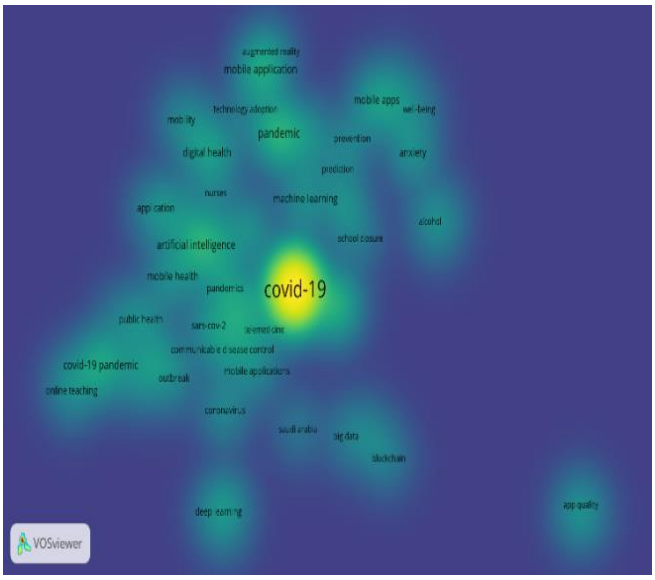


Fig. 9. Visualization of Density of Bibliometric Analysis.

In Fig. 9 it can be observed that Covid-19 is the most important term co-occurrence of all cited articles, such term co-occurrence is aligned to the present literature review. It is also observed that artificial intelligence and mobile application are the digital technologies with the highest term co-occurrence in relation to all cited articles.

After performing the bibliometric analysis, sixty-three (63) items were generated, grouped into ten (10) clusters shown in Table II. Finally, the clusters and their respective items were identified using the network visualization of the bibliometric analysis (Fig. 7), segmenting the clusters by their specific colors:

- Cluster 1 is colored magenta, "artificial intelligence".
- Cluster 2 is green, "application".
- Cluster 3 is colored blue, "pandemics".
- Cluster 4 is mustard-colored, "mental health".
- Cluster 5 is colored purple, "covid-19".
- Cluster 6 is light blue, "mhealth".
- Cluster 7 is colored orange, "mobile app".
- Cluster 8 is salmon-colored, "coronavirus".
- Cluster 9 is colored fuchsia, "machine learning".
- Cluster 10 is colored pink, "app quality".

TABLE II. ARTICLES, LINKS, TOTAL LINK STRENGTH, OCCURRENCES AND AVERAGE YEAR OF PUBLICATION

Article	Links	Total link strength	Occurrence	Avg. Pub. Year
Cluster 1				
artificial intelligence	12	23	10	2021.20
continuance intention	1	1	2	2022.00
covid-19 pandemic	10	10	8	2021.12
education	4	4	2	2020.50
intention to use	3	3	2	2021.00
internet	4	4	2	2021.00
online teaching	3	3	3	2020.33
outbreak	4	5	3	2021.33
prisma	3	3	2	2015.50
public health	6	8	4	2021.75
technology	8	8	3	2020.33
telehealth	4	4	2	2021.50
Cluster 2				
app	11	16	5	2021.00
application	4	6	3	2020.33
covid 19	4	4	3	2020.66
digital health	10	15	6	2021.33
intervention	5	7	2	2021.50
mobile	4	5	2	2020.50
mobile health	9	12	5	2021.20
mobility	4	6	3	2021.33
nurses	6	6	2	2021.50
protocol	6	7	2	2021.00
Cluster 3				
communicable disease control	8	12	3	2021.33
contact tracing	11	16	6	2020.50
humans	8	13	3	2021.00
mobile applications	9	12	3	2021.33
pandemics	9	12	4	2021.50
pneumonia	7	7	2	2021.00
sars-cov-2	10	14	4	2021.50
telemedicine	7	8	2	2021.50
Cluster 4				
anxiety	11	19	5	2021.20
depression	7	8	3	2020.66
mental health	9	12	5	2021.00
mobile apps	6	10	6	2021.33
prevention	6	7	3	2021.00
stress	4	5	2	2021.00
well-being	3	3	2	2021.00
Cluster 5				
big data	5	6	2	2021.00
blockchain	2	3	2	2021.00
covid-19	55	145	83	2021.10
diabetes	4	5	2	2020.50
health	5	6	2	2021.00
saudi arabia	4	5	2	2021.50
Cluster 6				
augmented reality	3	3	2	2021.50
mhealth	12	15	5	2021.20
mobile application	4	8	9	2021.22
pandemic	12	24	12	2020.75
technology adoption	4	4	2	2021.50
Cluster 7				
alcohol	5	10	2	2021.00
mindfulness	5	10	2	2021.00
mobile app	16	29	11	2020.72

Article	Links	Total link strength	Occurrence	Avg. Pub. Year
sleep	5	10	2	2021.00
Cluster 8				
convolutional neural networks	3	5	2	2020.50
coronavirus	9	11	4	2021.50
deep learning	4	7	3	2021.00
internet of things	4	4	3	2020.66
Cluster 9				
machine learning	6	11	5	2020.80
online learning	6	7	5	2021.20
prediction	5	6	2	2021.00
school closure	3	4	2	2021.00
Cluster 10				
app quality	3	6	2	2022.00
functional features	3	6	2	2022.00
mobile app rating scale	3	6	2	2022.00

Fig. 10 shows the number of selected articles grouped by database and digital technologies.

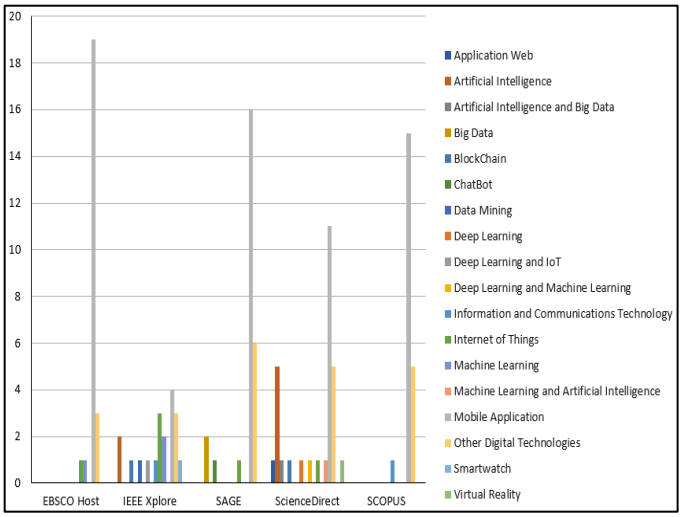


Fig. 10. Articles by Database and Digital Technologies.

Table III shows the digital technologies of the articles according to the results found.

TABLE III. CLASSIFICATION OF ARTICLES ACCORDING TO THE RESULTS OBTAINED

DIGITAL TECHNOLOGIES	ARTÍCLE
Application Web	[11]
Artificial Intelligence	[12]–[18]
Artificial Intelligence and Big Data	[19]
Big Data	[20], [21]
BlockChain	[22], [23]
ChatBot	[24]
Data Mining	[25]
Deep Learning	[26]
Deep Learning and IoT	[27]
Deep Learning and Machine Learning	[28]
Other Digital Technologies	[29]–[50]
Information and Communications Technology	[51], [52]
Internet of Things	[53]–[59]
Machine Learning	[60]–[62]
Machine Learning and Artificial Intelligence	[63]
Mobile Application	[64]–[126]
Smartwatch	[127]
Virtual Reality	[128]

Table IV shows the classification of articles according to the results found.

TABLE IV. CLASSIFICATION OF ITEMS ACCORDING TO CATEGORIES AND TOPICS OF FUNCTIONALITY

THEMES AND FUNCTIONALITY	REFERENCES
This article argues that the web pages allow informing about Covid-19, making it possible for the reader to become duly aware in order to have better sanitary control against Covid-19.	[11]
These articles argue that artificial intelligence allows a better study of the situation in order to achieve better health control against Covid-19.	[12]–[18]
This article argues that the use of artificial intelligence and Big Data make it possible to analyze a large amount of data in order to predict future Covid-19 scenarios and patterns, thus obtaining better health control against Covid-19.	[19]
These articles argue that the use of Big Data makes it possible to analyze a large amount of data in order to make a decision about Covid-19 and obtain better sanitary control against Covid-19.	[20], [21]
These articles argue that the use of the Block Chain makes it possible to share immutable data from medical research against Covid-19 and also to avoid misinformation, achieving better health surveillance against Covid-19 based on accurate information.	[22], [23]
This article argues that the use of ChatBots makes it possible to diagnose Covid-19 based on already defined questions, also, they can fulfill the role of informing to obtain better health control against Covid-19, based on truthful information.	[24]
This article argues that the use of data mining provides accurate information about a query, helps to have real statistics about Covid-19 allowing better health control Covid-19.	[25]
This article argues that the continuous use of Deep Learning makes it possible to refine it to anticipate responses and/or actions in certain scenarios in which Covid-19 is simulated, thus achieving better health control.	[26]
This article argues that the combination of Deep Learning with IoT opens up the possibility of creating intelligent objects that are refined according to their use, the IoT would be responsible for storing data while Deep Learning interprets it, obtaining tools that help health control against Covid-19.	[27]
This article argues that the combination of Deep Learning with Machine Learning opens the way to more accurate results as the algorithm to be used will adapt and learn to give us results that will help us to have better health control against Covid-19.	[28]
These articles argue that effective health control against Covid-19 requires the use of digital technologies.	[29]–[50]

THEMES AND FUNCTIONALITY	REFERENCES
These articles argue that ICTs play a very important role in helping us to have quick and easy access to global information and in times of pandemics they were of great help in promoting better health control against Covid-19.	[51], [52]
These articles argue that the use of the internet of things allows the development of new technologies, and, also, to rely on them, opening the way to telemedicine in times of pandemic, and obtaining better health control against Covid-19.	[53]–[58]
These articles argue that the application of Machine Learning, in the context of the topic, makes it possible to find patterns of Covid-19 infections and thus predict them in order to have better health control against Covid-19.	[60]–[62]
This article argues that the combination of Machine Learning and artificial intelligence opens the way to machines capable of learning on their own, in the present context, making possible the creation of robots that serve to monitor patients with Covid-19, thus avoiding contact with other human and the risk of contagion, achieving a better health control against Covid-19.	[63]
These articles argue that mobile applications are very useful because having an application related to Covid-19 allows us to be informed, consult and even monitor to achieve better sanitary control against Covid-19.	[64]–[126]
This article argues that the Smartwatch enables the detection of Covid-19 in the wearer as it can monitor heart rate and body temperature for diagnosis, thus achieving better health control of Covid-19.	[127]
This article argues that the use of virtual reality opens the way to being able to simulate Covid-19 patients and even scenarios to evaluate future decision-making in order to obtain a better health management of Covid-19.	[128]

Fig. 11 shows the number of selected items grouped by database and parameter.

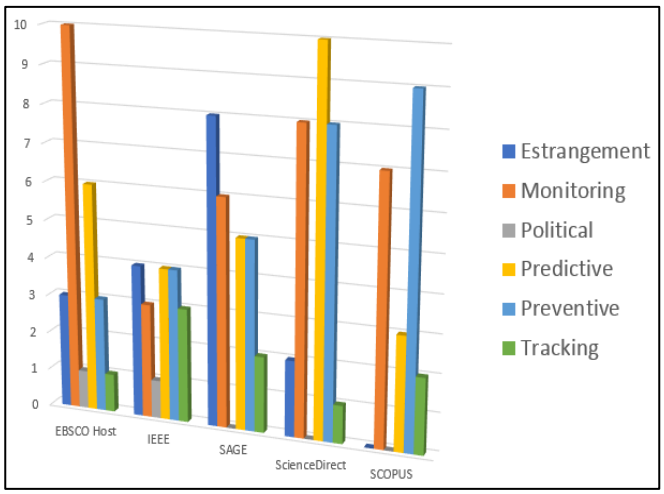


Fig. 11. Articles by Database and Parameter.

Table V shows the parameters of the articles according to the parameters found.

TABLE V. CLASSIFICATION OF ITEMS ACCORDING TO THE PARAMETERS OBTAINED

PARAMETERS	ARTÍCLES
Estrangement	[29]–[32], [51], [53], [54], [64]–[72], [127]
Monitoring	[12], [24], [27], [33]–[35], [52], [73]–[88], [90], [92]–[97], [129]–[132]
Political	[13], [98]
Predictive	[14]–[17], [19], [25], [26], [37]–[41], [60], [61], [63], [99]–[102], [104]–[109], [128], [133]
Preventive	[11], [20]–[22], [28], [42]–[47], [56]–[58], [110]–[122], [134]
Tracking	[18], [23], [49], [50], [123]–[126], [135]

Table VI shows the classification of articles according to the results found.

TABLE VI. CLASSIFICATION OF ITEMS ACCORDING TO PARAMETERS AND FUNCTIONALITY ISSUES

THEMES AND FUNCTIONALITY	REFERENCES
These articles take into account the parameter of distancing as they focus on creating strategies, and making use of technology, for the prevention of Covid-19, in scenarios where there is no rapprochement of individuals.	[29]–[32], [51], [53], [54], [64]–[72], [127]
These articles take into account the monitoring parameter as they focus on creating mobile applications and strategies based on Machine Learning, Deep Learning, and Artificial Intelligence since they can be monitoring an area where there was Covid-19 and even monitor the relationship between individuals and Covid-19.	[12], [24], [27], [33]–[35], [52], [73]–[88], [90], [92]–[97], [129]–[132]
These articles take into account the policy parameter as they encourage the use of technological tools in daily life and work so that, in times of pandemic, we adapt to study and work virtually, they also argue to consider very seriously the measures demanded by governments, to prevent Covid-19.	[13], [98]
These articles take into account the prediction parameter as they analyze data and facts, with the help of technological tools and Artificial Intelligence, making it possible for them to predict where Covid-19 could reemerge or if a community is about to suffer from it.	[14]–[17], [19], [25], [26], [37]–[41], [60], [61], [63], [99]–[102], [104]–[109], [128], [133]
These articles take into account the prevention parameter since they postulate measures, with the help of technological tools and Artificial Intelligence focusing on events that have already occurred, against Covid-19 to reduce contagion or a new outbreak.	[11], [20]–[22], [28], [42]–[47], [56]–[58], [110]–[122], [134]
These articles take into account the traceability parameter since they are based on following people or societies that have already suffered from Covid-1toto collect data and avoid possible Covid-19 infections.	[18], [23], [49], [50], [123]–[126], [135]

Fig. 12 shows the number of selected items grouped by database and quality attributes.

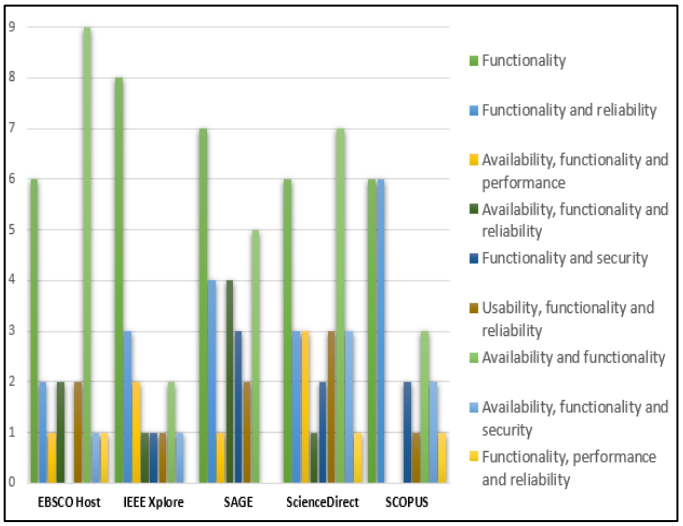


Fig. 12. Items by Database and Quality Attributes.

Table VII shows the quality attributes of the items according to the quality attributes found.

TABLE VII. CLASSIFICATION OF ARTICLES ACCORDING TO THE QUALITY ATTRIBUTES OBTAINED

QUALITY ATTRIBUTES	ARTÍCLES
Availability and functionality	[12], [25], [30], [31], [34], [35], [38], [48], [56], [63], [69], [74], [77], [79], [82], [88], [92], [105], [112], [115], [117], [120], [128], [133], [136]
Availability, functionality, and reliability	[24], [40], [66], [73], [80], [87], [100], [114]
Availability, functionality, and performance	[11], [16], [29], [46], [62], [86], [110]
Availability, functionality, and security	[37], [44], [58], [64], [95], [122], [137]
Functionality	[14], [15], [23], [27], [28], [32], [36], [39], [43], [45], [47], [49], [51], [55], [57], [60], [61], [65], [71], [72], [75], [81], [83], [90], [97], [101], [102], [108], [116], [119], [124], [138], [139]
Functionality and reliability	[17], [19], [20], [26], [41], [52], [53], [67], [68], [76], [94], [96], [99], [104], [109], [121], [123], [127]
Functionality and security	[18], [21], [22], [50], [54], [70], [125], [126]
Functionality, performance, and reliability	[107], [113]
Usability, functionality, and reliability	[33], [78], [84], [85], [89], [93], [106], [111], [118]

Table VIII shows the classification of articles according to the results found.

TABLE VIII. CLASSIFICATION OF ITEMS ACCORDING TO QUALITY ATTRIBUTES AND FUNCTIONALITY ISSUES

THEMES AND FUNCTIONALITY	REFERENCES
These articles take into account the quality attributes of availability and functionality, mentioning that these quality attributes are the priorities for the development of effective and viable software for sanitary control against Covid-19.	[12], [25], [30], [31], [34], [35], [38], [48], [56], [63], [69], [74], [77], [79], [82], [88], [92], [105], [112], [115], [117], [120], [128], [133], [136]
These articles take into account the quality attributes of availability, functionality and reliability, taking into account this, it will be possible to develop software for sanitary control against Covid-19 with high-quality standards.	[24], [40], [66], [73], [80], [87], [100], [114]
These articles take into account the quality attributes of availability, functionality and performance, which would serve to develop software with the necessary functions to solve the problems presented by Covid-19.	[11], [16], [29], [46], [62], [86], [110]
These articles take into account the quality attributes of availability, functionality, and security, since the development of software that is ready for use at any time, works as designed, and contains a level of resistance to be breached, would be completely viable to combat Covid-19.	[37], [44], [58], [64], [95], [122], [137]
These articles take into account the quality attribute of functionality since they consider that it is the main quality attribute for the software to be able to perform the functions for which it was programmed or developed, being highly effective in time of pandemic by Covid-19.	[14], [15], [23], [27], [28], [32], [36], [39], [43], [45], [47], [49], [51], [55], [57], [60], [61], [65], [71], [72], [75], [81], [83], [90], [97], [101], [102], [108], [116], [119], [124], [138], [139]
These articles take into account the quality attributes of functionality and reliability since a software must fulfill the functions for which it was created, but also that software must remain operational over time, these requirements are perfectly aligned to address the problems that Covid-19 is presenting.	[17], [19], [20], [26], [41], [52], [53], [67], [68], [76], [94], [96], [99], [104], [109], [121], [123], [127]
These articles take into account the quality attributes of functionality and security since they emphasize that all software must do the job for which it was developed and, nowadays, all software must have a decent degree of security since the Covid-19 pandemic has increased the rate of cybercrime.	[18], [21], [22], [50], [54], [70], [125], [126]
These articles take into account the quality attributes of functionality, performance, and reliability, they mention these quality attributes because they consider that software must perform its functions correctly with speed and accuracy, being able to operate with a large amount of data, such as the data generated by the Covid-19 pandemic, without having a long waiting time.	[107], [113]
These articles take into account the quality attributes of usability, functionality, and reliability, since a viable and effective software that is focused on the study of Covid-19, must be simple and understandable for all audiences, with correctly programmed and operational functions.	[33], [78], [84], [85], [89], [93], [106], [111], [118]

IV. DISCUSSION

This systematic literature review is intended to answer the following questions.

RQ1. What digital technologies allow for better control, follow-up, and monitoring of the health status of students, teachers, and staff in educational centers against Covid-19?

According to Fig. 10, it can be seen that the articles related to the present topic in question use the digital technologies of; Mobile Application, Artificial Intelligence, Other Digital Technologies, etc. This result indicates that these technological categories allow sanitary control against Covid-19.

According to Table III, and commenting on it in Table IV, it can be seen that the digital technologies related to the present topic use the digital technology of "Mobile Application". This result indicates that this digital technology is one of the most used in allowing to have sanitary control against Covid-19 and is aligned to today's technological era, where we all use a mobile device.

RQ2. What parameters should be taken into account to make effective sanitary control against Covid-19 in educational centers through the use of a mobile application?

According to Fig. 11, it can be seen that the articles related to this topic use the technology of distancing, monitoring, policy, prediction, prevention, and tracking. This result indicates that these parameters allow sanitary control against Covid-19. According to Table V, and commenting on it in Table VI, it can be seen that the parameters related to the present topic use the "Monitoring" parameter. This result indicates that this parameter is one of the most used in allowing sanitary control against Covid-19 since constant monitoring makes possible the collection of data and its subsequent analysis.

RQ3. What quality attributes must it contain for the viability of the mobile application for the implementation of a sanitary control against Covid-19 in educational centers?

According to Fig. 12, it can be seen that the articles related to the present topic use the quality attributes; functionality, availability and functionality, functionality and reliability, etc. This result indicates that these quality attributes allow the creation of an effective software directed to have sanitary control against Covid-19.

According to Table VII, and commenting on it in Table VIII, it can be seen that the parameters related to the present topic use the "functionality" parameter. This result indicates that this quality attribute is one of the most used in allowing sanitary control against Covid-19 since this attribute focuses on the ability of the system to perform the task for which it was developed.

RQ4. Which countries have the most research, in the last three years, related to health monitoring against Covid-19 in schools?

According to Fig. 5, it can be seen that the articles related to this topic come from the continents of Asia, America, and Europe (from highest to lowest). This result indicates that there is a greater knowledge of the technologies related to sanitary control of Covid-19.

According to Fig. 6, it can be seen that the articles related to this topic come mostly from the United States and China. This result indicates that there is more experience in sanitary control against Covid-19 in these countries.

V. Proposed Model

The following is a proposed model based on mobile applications for the implementation of sanitary control against Covid-19 in educational centers, aligned with the data collected from the articles related to the present topic (Fig. 13). The proposed model is related to the article [140]. In Fig. 12 we can see the proposed model, which includes the phases of the data set: It is the literature systematization of all articles found related to the topic and complying with the standards of the article inclusion chart (Fig. 1) and the inclusion and exclusion criteria (Table I). In the ideal characteristics of the mobile applications for the implementation of sanitary control against covid-19 in educational centers, three main characteristics have been taken into account, which are:

Additional digital technologies: After systematizing the literature, we were able to identify, according to Fig. 10 and commenting on it in Tables III and IV, the additional digital technologies that are most used at the time of the development of a mobile application for the implementation of a sanitary control against Covid-19. Beginning to order them by their level of use (except for mobile applications and digital technology), in the first instance we have artificial intelligence because the additional technology mentioned would allow a better study based on complex algorithms which would be interpreted by the machines making more effective the functions of the mobile application. In the second instance, we have the internet of things since the implementation of communication between the mobile application and the intelligent sensors of the mobile device would make the mobile application more complete and could use data provided by the user himself, such as his location via GPS his daily steps or speed of movement through the accelerometer.

Parameter: Another ideal characteristic for the development of an effective and viable mobile application is the parameter on which it will be based. In the present systematic review of the literature, the parameters on which the cited articles were based to obtain a better sanitary control against Covid-19 were identified according to Fig. 11 and commented in Table V and VI. It was evident that the most used parameter was monitoring, since it allows collecting, studying, and analyzing the information obtained for a constant follow-up, i.e. it will be possible to analyze the information of patients or communities where there was Covid-19 to identify symptoms of the virus, precedents and even to identify where it could outbreak, Following this parameter, it would be possible to have a better sanitary control against Covid-19 since it would be evaluating, in every certain period, the health avoiding contagions, identifying possible resurgence and possible people prone to contract Covid-19, currently, the mobile device would be aligned to support itself in the monitoring to give more accurate results of an individual diagnosis and that would be thanks to the sensors contained in each mobile device. The second most used parameter is preventive since its main objective is to reduce future Covid-19 infections, that is to say,

to study real, truthful, and reliable data so that later they can carry out a better control against Covid-19, besides, with the help of technology and the creation of software and apps, more accurate data could be obtained to identify the infected areas, obtaining, as a result, an alert in that area, achieving to reduce infections or a new outbreak.

Quality attributes: Finally, we have identified a third and last ideal characteristic. When the systematization of the literature was done, quality attributes were identified to reach the development of a mobile application for the implementation of sanitary control against Covid-19. According to Fig. 12 and comments in Tables VII and VIII, the most frequent quality attribute is the functionality, since it allows the system to perform the work for which it was created, that is, to give us accurate results of the operations we perform, so that when we want to consult the Covid-19 infection rate or perform a diagnosis of Covid-19 using the software, Another example is when we use a mobile application that by me using bits, our daily route and monitoring it by means of using us with certain information for which it was developed and thus fulfills its functionality. The second quality attributes were availability and functionality, previously we have already explained the functionality so we would complement the explanation now with reference to concerning, then when you mention availability, we refer to the ability of all software to be executed when the user needs it, i.e. it must be accessible and usable.

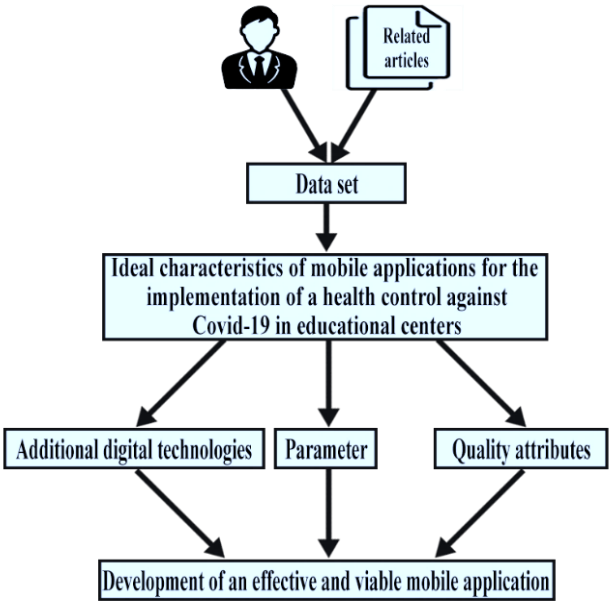


Fig. 13. Proposal Template.

Development of an effective and viable mobile application: With the three ideal characteristics, identified at the time when the systematization of the literature was performed, this point could be fulfilled, because while a mobile application contains additional digital technologies then the final product will be more complete, will contain many functions, will obtain more precision in its results or final functions. Do not forget that the parameter on which the mobile application will be based can get improved and be clearer about the functions that we would like to fulfill for the mobile application to develop. Finally, the quality attributes are very essential in all software because they are indicators that reflect how well the system meets the needs of stakeholders.

VI. Related Articles

Other systematic literature review studies conducted such as [141] based on the identification of the factor affecting the intention of continuous use of the mHealth application, compared to the present systematic review, they identified 354 articles which by removing the irrelevant and duplicated ones, 25 selected articles were left for the systematization of the literature. From this, they identified the factors for the continued use of mHealth applications (identifying a total of 39) some factors are satisfaction, quality of service, monetary cost, age, and education, among others. It concludes by identifying the five most frequent factors, which are satisfaction, the usefulness of the mHealth, quality service, training of the service, and ease of use of the service. It shows that users prefer mHealth to meet this expectation, since a mobile application should not only be easy to use but also have quality service and trust in it.

On the other hand, the following systematic review of the literature [142], searched for articles in the PubMed and Scopus databases to identify mHealth applications used for the prevention, treatment, or management of COVID-19. They identified a total of 728 articles of which, using the PRISMA methodology, they were left with 12 articles for the systematization of the literature. The author concluded that the studies in the articles he reviewed were not of high quality since Covid-19 had to generate responses and very premature development of digital tools for health by the scientific community. He emphasizes that a more longitudinal study with a rigorous design is required for a better evaluation of mobile applications against Covid-19.

Finally, in the following systematic review of the literature, [143] based on the study of data privacy during pandemics: a systematic literature review of smartphone applications Covid-19, identified 808 articles, using Liao's methodology, they were left with 35 articles for systematization of the literature. It relates that data privacy or information privacy often revolves around whether or not the data stored in the mobile application is shared with third parties. He also emphasizes that there are security policies for users and this is very important because in pandemic times there was a high increase in the use of mobile applications, especially against Covid-19, so this point of data protection is vital for the current technological era.

VII. Conclusion

After having carried out systematic literature research of 119 articles related to the topic in question, it is concluded that:

The digital tools or technologies that allow better control, follow-up, and monitoring against Covid-19 of the health status of students, teachers, and staff in educational centers are the digital technologies associated with "Mobile application", "Digital technology" and "Artificial intelligence". Likewise, most of the authors, from the articles reviewed, choose to base their article on a parameter focused on monitoring Covid-19.

Therefore, it was evident that most authors rely on the parameter "Monitoring" because they can observe and study patients who have contracted Covid-,19 and even take a study of the locality where there was Covid-19 which would help them to obtain updated reports and conduct accurate research.

Regarding the quality attributes that must be contained for the viability of the mobile application for the implementation of a sanitary control against Covid-19 in educational centers, they are availability, and functionality since this allows the system to perform the work for which it was created and to be available for use, meeting the needs of users. It was also concluded that the countries with the most research, in the last three years, related to sanitary control against Covid-19 in educational centers are the United States and China, showing that these countries have greater experience in sanitary control against Covid-19.

Finally, a proposed model was postulated to achieve the development of an effective and viable mobile application based on the three ideal characteristics previously explained. This systematic review can also be useful for use in future research on digital technologies, parameters, and quality attributes for the implementation of a health control against Covid-19 in educational centers, as well as identifying the countries that have more experience in this subject. In contrast, we can also rescue from the article [140] a graphic suggestion aligned to our conclusion (Fig. 14):

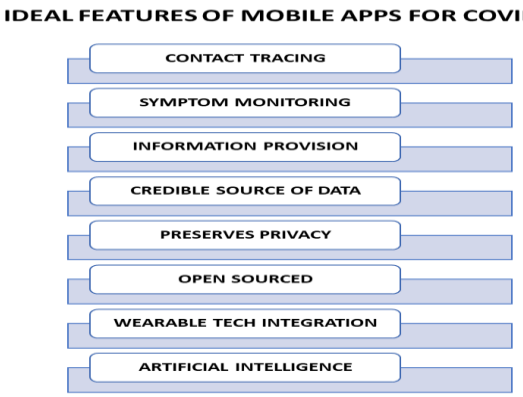


Fig. 14. Ideal Features of Mobile Apps for COVID-19.

Fig. 13 suggests some features that should be available in mobile applications for Covid-19. Other important functionalities that can be integrated into these contact-tracking apps include features for automatic symptom monitoring and information provision. The addition of these features will provide a more holistic public health approach in response to the situation. As technology advances, the symptom tracking algorithm can be enhanced and adapted to the pandemic to improve its diagnostic accuracy. Wearable devices, such as smartwatches and smart bracelets, will become more common and integrated into everyday life; therefore, these can potentially assist in vital monitoring of the health status of vulnerable populations. Through machine learning and artificial intelligence methods, automatic and rapid identification of suspicious infections will be more accurate in the future.

References

[1] D. Dong et al., "The Role of Imaging in the Detection and Management of COVID-19: A Review," IEEE Reviews in Biomedical Engineering, vol. 14, pp. 16–29, 2021, doi: 10.1109/RBME.2020.2990959.

[2] S. Gupta, D. Rastogi, K. Chauhan, and S. Sharma, "Comparative Analysis of the 1stand 2ndWave of COVID-19 and Visualizing the Increasing and Decreasing of COVID-19," Proceedings of the 2021 10th International Conference on System Modeling and Advancement in Research Trends, SMART 2021, pp. 324–329, 2021, doi: 10.1109/SMART52563.2021.9676279.

[3] A. Khattar, P. R. Jain, and S. M. K. Quadri, "Effects of the Disastrous Pandemic COVID 19 on Learning Styles, Activities and Mental Health of Young Indian Students-A Machine Learning Approach," Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020, pp. 1190–1195, May 2020, doi: 10.1109/ICICCS48265.2020.9120955.

[4] A. Alalawi, "A Survey on E-learning Methods and Effectiveness in Public Bahrain Schools during the COVID-19 pandemic," 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2021, pp. 574–579, Sep. 2021, doi: 10.1109/3ICT53449.2021.9582021.

[5] "View of Teachers' opinions on (urgent) distance education activities during the pandemic period." https://www.syncsci.com/journal/AMLER/article/view/AMLER.2022.02.005/658 (accessed Aug. 20, 2022).

[6] R. I. Sifat, M. M. Ruponty, Md. K. Rahim Shuvo, M. Chowdhury, and S. M. Suha, "Impact of COVID-19 pandemic on the mental health of school-going adolescents: insights from Dhaka city, Bangladesh," Heliyon, vol. 8, no. 4, p. e09223, Apr. 2022, doi: 10.1016/J.HELIYON.2022.E09223.

[7] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," Systematic Reviews, vol. 10, no. 1, pp. 1–11, Dec. 2021, doi: 10.1186/S13643-021-01626-4/FIGURES/1.

[8] P. Ahmad, J. A. Asif, M. K. Alam, and J. Slots, "A bibliometric analysis of Periodontology 2000," Periodontol 2000, vol. 82, no. 1, pp. 286–297, Feb. 2020, doi: 10.1111/PRD.12328.

[9] A. Tasdelen and A. R. Ugur, "Artificial Intelligence Research on COVID-19 Pandemic: A Bibliometric Analysis," ISMSIT 2021 - 5th International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings, pp. 693–699, 2021, doi: 10.1109/ISMSIT52890.2021.9604573.

[10] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," Scientometrics, vol. 84, no. 2, pp. 523–538, 2010, doi: 10.1007/s11192-009-0146-3.

[11] I. Carvalho et al., "COVID-19 BR: A web portal for COVID-19 information in Brazil," Procedia Computer Science, vol. 196, pp. 525–532, Jan. 2022, doi: 10.1016/J.PROCS.2021.12.045.

[12] M. Hajder, P. Hajder, T. Gil, M. Krzywda, J. Kolbusz, and M. Liput, "Architecture and organization of a Platform for diagnostics, therapy and post-covid complications using AI and mobile monitoring," Procedia Computer Science, vol. 192, pp. 3711–3721, Jan. 2021, doi: 10.1016/J.PROCS.2021.09.145.

[13] O. Tutsoy, "COVID-19 Epidemic and Opening of the Schools: Artificial Intelligence-Based Long-Term Adaptive Policy Making to Control the Pandemic Diseases," IEEE Access, vol. 9, pp. 68461–68471, 2021, doi: 10.1109/ACCESS.2021.3078080.

[14] F. A. Muqtadiroh et al., "Fuzzy Unsupervised Approaches to Analyze Covid-19 Spread for School Reopening Decision Making," IECON Proceedings (Industrial Electronics Conference), vol. 2021-October, Oct. 2021, doi: 10.1109/IECON48115.2021.9589699.

[15] J. Aljizawi, D. Dalloul, L. Ghryani, S. Aldabbagh, and T. Brahimi, "A Survey of Artificial Intelligence Solutions in Response to the COVID-19 Pandemic in Saudi Arabia," Procedia Computer Science, vol. 194, pp. 190–201, Jan. 2021, doi: 10.1016/J.PROCS.2021.10.073.

[16] M. Bhatia, A. Manocha, T. A. Ahanger, and A. Alqahtani, "Artificial intelligence-inspired comprehensive framework for Covid-19 outbreak control," Artificial Intelligence in Medicine, vol. 127, p. 102288, May 2022, doi: 10.1016/J.ARTMED.2022.102288.

[17] S. Sarker, L. Jamal, S. F. Ahmed, and N. Irtisam, "Robotics and artificial intelligence in healthcare during COVID-19 pandemic: A systematic review," Robotics and Autonomous Systems, vol. 146, p. 103902, Dec. 2021, doi: 10.1016/J.ROBOT.2021.103902.

[18] H. Haneya, D. Alkaf, F. Bajammal, and T. Brahimi, "A Meta-Analysis of Artificial Intelligence Applications for Tracking COVID-19: The Case of the U.A.E.," Procedia Computer Science, vol. 194, pp. 180–189, Jan. 2021, doi: 10.1016/J.PROCS.2021.10.072.

[19] P. Galetsi, K. Katsaliaki, and S. Kumar, "The medical and societal impact of big data analytics and artificial intelligence applications in combating pandemics: A review focused on Covid-19," Social Science & Medicine, vol. 301, p. 114973, May 2022, doi: 10.1016/J.SOCSCIMED.2022.114973.

[20] A. Poom, O. Järv, M. Zook, and T. Toivonen, "COVID-19 is spatial: Ensuring that mobile Big Data is used for social good.," Big Data Soc, vol. 7, no. 2, p. 2053951720952088, Jul. 2020, doi: 10.1177/2053951720952088.

[21] Y. E. Park, "Developing a COVID-19 Crisis Management Strategy Using News Media and Social Media in Big Data Analytics," https://doi.org/10.1177/08944393211007314, Apr. 2021, doi: 10.1177/08944393211007314.

[22] W. Y. Ng et al., "Blockchain applications in health care for COVID-19 and beyond: a systematic review," The Lancet Digital Health, vol. 3, no. 12, pp. e819–e829, Dec. 2021, doi: 10.1016/S2589-7500(21)00210-7.

[23] S. Peng, L. Bai, L. Xiong, Q. Qu, X. Xie, and S. Wang, "GeoAI-based epidemic control with geo-social data sharing on blockchain," 2020 IEEE International Conference on E-Health Networking, Application and Services, HEALTHCOM 2020, Mar. 2021, doi: 10.1109/HEALTHCOM49281.2021.9399031.

[24] Y. Zhu, R. Wang, and C. Pu, "'I am chatbot, your virtual mental health adviser.' What drives citizens' satisfaction and continuance intention toward mental health chatbots during the COVID-19 pandemic? An empirical study in China," https://doi.org/10.1177/20552076221090031, vol. 8, p. 205520762210900, Mar. 2022, doi: 10.1177/20552076221090031.

[25] A. Mohammed, A. Khedr, D. Alhaj, R. Al Khalifa, and A. M. Zeki, "The Impact of Family, Lifestyle, and COVID-19 Factors on Private High School Students' Academic Performance: Data Mining Approach," 2021 International Conference on Decision Aid Sciences and Application, DASA 2021, pp. 309–313, 2021, doi: 10.1109/DASA53625.2021.9682282.

[26] S. Serte and H. Demirel, "Deep learning for diagnosis of COVID-19 using 3D CT scans," Computers in Biology and Medicine, vol. 132, p. 104306, May 2021, doi: 10.1016/J.COMPBIOMED.2021.104306.

[27] N. A. Othman, M. Z. N. Al-Dabagh, and I. Aydin, "A New Embedded Surveillance System for Reducing COVID-19 Outbreak in Elderly Based on Deep Learning and IoT," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI 2020, Oct. 2020, doi: 10.1109/ICDABI51230.2020.9325651.

[28] E. Mbunge, S. Simelane, S. G. Fashoto, B. Akinnuwesi, and A. S. Metfula, "Application of deep learning and machine learning models to detect COVID-19 face masks - A review," Sustainable Operations and Computers, vol. 2, pp. 235–245, Jan. 2021, doi: 10.1016/J.SUSOC.2021.08.001.

[29] B. Shambare and C. Simuja, "A Critical Review of Teaching With Virtual Lab: A Panacea to Challenges of Conducting Practical Experiments in Science Subjects Beyond the COVID-19 Pandemic in Rural Schools in South Africa," Journal of Educational Technology Systems, vol. 50, no. 3, pp. 393–408, Mar. 2022, doi: 10.1177/00472395211058051.

[30] M. A. Khan, Vivek, M. K. Nabi, M. Khojah, and M. Tahir, "Students' perception towards e-learning during covid-19 pandemic in India: An empirical study," Sustainability (Switzerland), vol. 13, no. 1, pp. 1–14, Jan. 2021, doi: 10.3390/SU13010057.

[31] A. Asadpour, "Student challenges in online architectural design courses in Iran during the COVID-19 pandemic," https://doi.org/10.1177/20427530211022923, vol. 18, no. 6, pp. 511–529, Jun. 2021, doi: 10.1177/20427530211022923.

[32] A. Jakoet-Salie and K. Ramalobe, "The digitalization of learning and teaching practices in higher education institutions during the Covid-19 pandemic," https://doi.org/10.1177/01447394221092275, p. 014473942210922, Apr. 2022, doi: 10.1177/01447394221092275.

[33] N. A. Muhammad, "The Usability and Feasibility of DailyCalm Application in Reducing Stress among Adolescents During COVID-19 Pandemic," Medicine & Health, vol. 16, no. 2, pp. 216–226, Dec. 2021, doi: 10.17576/MH.2021.1602.16.

[34] E. Mbunge, J. Batani, G. Gaobotse, and B. Muchemwa, "Virtual healthcare services and digital health technologies deployed during coronavirus disease 2019 (COVID-19) pandemic in South Africa: a systematic review," Global Health Journal, Mar. 2022, doi: 10.1016/J.GLOHJ.2022.03.001.

[35] P. Wantanakorn, "Digital media and child development in the covid-19 pandemic: Benefits, disadvantages, and effective approaches," Journal of the Medical Association of Thailand, vol. 104, no. 9, pp. 1563–1569, Sep. 2021, doi: 10.35755/JMEDASSOCTHAI.2021.09.12543.

[36] G. L. Vasconcelos, G. C. Duarte-Filho, A. A. Brum, R. Ospina, F. A. G. Almeida, and A. M. S. Macêdo, "Situation of COVID-19 in Brazil in August 2020: An Analysis via Growth Models as Implemented in the ModInterv System for Monitoring the Pandemic," Journal of Control, Automation and Electrical Systems, vol. 33, no. 2, pp. 645–663, Apr. 2022, doi: 10.1007/s40313-021-00853-3.

[37] Y. Martínez-Díaz, H. Méndez-Vázquez, L. S. Luevano, M. Nicolás-Díaz, L. Chang, and M. González-Mendoza, "Towards Accurate and Lightweight Masked Face Recognition: An Experimental Evaluation," IEEE Access, vol. 10, pp. 7341–7353, 2022, doi: 10.1109/ACCESS.2021.3135255.

[38] M. Chakraborty, M. S. Mahmud, T. J. Gates, and S. Sinha, "Analysis and Prediction of Human Mobility in the United States during the Early Stages of the COVID-19 Pandemic using Regularized Linear Models," https://doi.org/10.1177/03611981211067794, p. 036119812110677, Jan. 2022, doi: 10.1177/03611981211067794.

[39] C. J. H. Kim and A. M. Padilla, "Technology for Educational Purposes Among Low-Income Latino Children Living in a Mobile Park in Silicon Valley: A Case Study Before and During COVID-19," https://doi.org/10.1177/0739986320959764, vol. 42, no. 4, pp. 497–514, Sep. 2020, doi: 10.1177/0739986320959764.

[40] A. Khan et al., "A combined model for COVID-19 pandemic control: The application of Haddon's matrix and community risk reduction tools combined," Journal of Infection and Public Health, vol. 15, no. 2, pp. 261–269, Feb. 2022, doi: 10.1016/J.JIPH.2022.01.006.

[41] S. A. Müller et al., "Predicting the effects of COVID-19 related interventions in urban settings by combining activity-based modelling, agent-based simulation, and mobile phone data," PLoS ONE, vol. 16, no. 10 October, Oct. 2021, doi: 10.1371/JOURNAL.PONE.0259037.

[42] E. Jordan, D. E. Shin, S. Leekha, and S. Azarm, "Optimization in the Context of COVID-19 Prediction and Control: A Literature Review," IEEE Access, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3113812.

[43] M. Pears, M. Yiasemidou, M. A. Ismail, D. Veneziano, and C. S. Biyani, "Role of immersive technologies in healthcare education during the COVID-19 epidemic," Scottish Medical Journal, vol. 65, no. 4, pp. 112–119, Nov. 2020, doi: 10.1177/0036933020956317.

[44] K. Intawong, D. Olson, and S. Chariyalertsak, "Application technology to fight the COVID-19 pandemic: Lessons learned in Thailand," Biochemical and Biophysical Research Communications, vol. 538, pp. 231–237, Jan. 2021, doi: 10.1016/J.BBRC.2021.01.093.

[45] Y. Wang et al., "Applications of additive manufacturing (AM) in sustainable energy generation and battle against COVID-19 pandemic: The knowledge evolution of 3D printing," Journal of Manufacturing Systems, vol. 60, pp. 709–733, Jul. 2021, doi: 10.1016/J.JMSY.2021.07.023.

[46] S. Whitelaw, M. A. Mamas, E. Topol, and H. G. C. Van Spall, "Applications of digital technology in COVID-19 pandemic planning and response," The Lancet Digital Health, vol. 2, no. 8, pp. e435–e440, Aug. 2020, doi: 10.1016/S2589-7500(20)30142-4.

[47] L. Birrell, A. Furneaux-Bate, C. Chapman, and N. C. Newton, "A mobile peer intervention for preventing mental health and substance use problems in adolescents: Protocol for a randomized controlled trial (the

mind your mate study)," JMIR Research Protocols, vol. 10, no. 7, Jul. 2021, doi: 10.2196/26796.

[48] H. C. Sun, X. F. Liu, Z. W. Du, X. K. Xu, and Y. Wu, "Mitigating COVID-19 Transmission in Schools with Digital Contact Tracing," IEEE Transactions on Computational Social Systems, vol. 8, no. 6, pp. 1302–1310, Dec. 2021, doi: 10.1109/TCSS.2021.3073109.

[49] N. Kalyanaraman and M. R. Fraser, "Containing COVID-19 Through Contact Tracing: A Local Health Agency Approach," Public Health Reports, vol. 136, no. 1, pp. 32–38, Jan. 2021, doi: 10.1177/0033354920967910.

[50] T. Cinque, "Protecting communities during the COVID-19 global health crisis: health data research and the international use of contact tracing technologies," Humanities and Social Sciences Communications, vol. 9, no. 1, Dec. 2022, doi: 10.1057/s41599-022-01078-8.

[51] A. Spernjak, "Using ICT to Teach Effectively at COVID-19," 2021 44th International Convention on Information, Communication and Electronic Technology, MIPRO 2021 - Proceedings, pp. 617–620, 2021, doi: 10.23919/MIPRO52101.2021.9596878.

[52] M. Bortoluzzi, T. M. Sgaramella, L. Ferrari, V. Drąsutė, and V. Šarauskytė, "Building Emotionally Stable, Inclusive, and Healthy Communities with ICT: From State of the Art to PSsmile App," Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, vol. 401 LNICST, pp. 163–178, 2021, doi: 10.1007/978-3-030-91421-9_13.

[53] T. Bhowmik, R. Mojumder, I. Banerjee, G. Das, and A. Bhattacharya, "IoT Based Non-Contact Portable Thermal Scanner for COVID Patient Screening," 2020 IEEE 17th India Council International Conference, INDICON 2020, Dec. 2020, doi: 10.1109/INDICON49873.2020.9342203.

[54] P. V. Bindu, K. D. Al-Hanawi, A. M. Al-Abri, and V. Mahadevan, "IoT based safety system for school children: A contactless access control for post covid school conveyance," 2021 2nd International Conference for Emerging Technology, INCET 2021, May 2021, doi: 10.1109/INCET51464.2021.9456314.

[55] S. S. Vedaei et al., "COVID-SAFE: An IoT-based system for automated health monitoring and surveillance in post-pandemic life," IEEE Access, vol. 8, pp. 188538–188551, 2020, doi: 10.1109/ACCESS.2020.3030194.

[56] Z. Chen, S. Khan, M. Abbas, S. Nazir, and K. Ullah, "Enhancing Healthcare through Detection and Prevention of COVID-19 Using Internet of Things and Mobile Application," Mobile Information Systems, vol. 2021, 2021, doi: 10.1155/2021/5291685.

[57] W. L. Lin, C. H. Hsieh, T. S. Chen, J. Chen, J. Le Lee, and W. C. Chen, "Apply IOT technology to practice a pandemic prevention body temperature measurement system: A case study of response measures for COVID-19," https://doi.org/10.1177/15501477211018126, vol. 17, no. 5, May 2021, doi: 10.1177/15501477211018126.

[58] A. H. Mohd Aman, W. H. Hassan, S. Sameen, Z. S. Attarbashi, M. Alizadeh, and L. A. Latiff, "IoMT amid COVID-19 pandemic: Application, architecture, technology, and security," Journal of Network and Computer Applications, vol. 174, p. 102886, Jan. 2021, doi: 10.1016/J.JNCA.2020.102886.

[59] J. Amachi-Choqque and M. Cabanillas-Carbonell, "IoT System for Vital Signs Monitoring in Suspicious Cases of Covid-19," International Journal of Advanced Computer Science and Applications, vol. 12, no. 2, pp. 174–180, 2021, doi: 10.14569/IJACSA.2021.0120223.

[60] F. Faisal, M. M. Nishat, M. A. Mahbub, M. M. I. Shawon, and M. M. U. H. Alvi, "Covid-19 and its impact on school closures: A predictive analysis using machine learning algorithms," 2021 International Conference on Science and Contemporary Technologies, ICSCT 2021, 2021, doi: 10.1109/ICSCT53883.2021.9642617.

[61] I. Akour, M. Alshurideh, B. Al Kurdi, A. Al Ali, and S. Salloum, "Using machine learning algorithms to predict people's intention to use mobile learning platforms during the COVID-19 pandemic: Machine learning approach," JMIR Medical Education, vol. 7, no. 1, Jan. 2021, doi: 10.2196/24032.

[62] J. Ispahany and R. Islam, "Detecting malicious COVID-19 URLs using machine learning techniques," 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events, PerCom Workshops 2021, pp. 718–723, Mar. 2021, doi: 10.1109/PERCOMWORKSHOPS51409.2021.9431064.

[63] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review," Chaos, Solitons & Fractals, vol. 139, p. 110059, Oct. 2020, doi: 10.1016/J.CHAOS.2020.110059.

[64] E. C. Anyanwu, R. Parker Ward, A. Shah, V. Arora, and C. A. Umscheid, "A mobile app to facilitate socially distanced hospital communication during COVID-19: Implementation experience," JMIR Mhealth Uhealth, vol. 9, no. 2, Feb. 2021, doi: 10.2196/24452.

[65] D. Sharma and M. Alam, "Aesthetics, Emotions, and the Use of Online Education Apps Post-COVID-19 Pandemic," https://doi.org/10.1177/21582440221093047, vol. 12, no. 2, p. 215824402210930, Apr. 2022, doi: 10.1177/21582440221093047.

[66] E. Krisiunas and L. Sibomana, "Benefits of Technology in the Age of COVID-19 and Diabetes..Mobile Phones From a Rwanda Perspective," Journal of Diabetes Science and Technology, vol. 14, no. 4, pp. 748–749, Jul. 2020, doi: 10.1177/1932296820930032.

[67] A. Q. Blebil et al., "Exploring the eHealth literacy and mobile health application utilisation amongst Malaysian pharmacy students," https://doi.org/10.1177/1357633X221077869, Feb. 2022, doi: 10.1177/1357633X221077869.

[68] K. Kaufmann, C. Peil, and T. Bork-Hüffer, "Producing In Situ Data From a Distance With Mobile Instant Messaging Interviews (MIMIs): Examples From the COVID-19 Pandemic," https://doi.org/10.1177/16094069211029697, vol. 20, Aug. 2021, doi: 10.1177/16094069211029697.

[69] S. Tan, "Remote learning through a mobile application in gifted education," https://doi.org/10.1177/02614294211069627, vol. 38, no. 1, pp. 95–114, Dec. 2021, doi: 10.1177/02614294211069627.

[70] S. Duguay, C. Dietzel, and D. Myles, "The year of the 'virtual date': Reimagining dating app affordances during the COVID-19 pandemic," https://doi.org/10.1177/14614448211072257, Jan. 2022, doi: 10.1177/14614448211072257.

[71] Y. Zhou, R. Xu, D. Hu, Y. Yue, Q. Li, and J. Xia, "Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data," The Lancet Digital Health, vol. 2, no. 8, pp. e417–e424, Aug. 2020, doi: 10.1016/S2589-7500(20)30165-5.

[72] Z. Liu, Q. He, X. Zhou, and Y. Hai, "Mobile application–based behaviour change techniques to encourage quarantine compliance during the COVID-19 pandemic," Public Health, vol. 197, pp. e6–e7, Aug. 2021, doi: 10.1016/J.PUHE.2020.11.017.

[73] Muladi et al., "Development of the Personnel Monitoring System Using Mobile Application and Real-Time Database during the COVID19 Pandemic," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020, pp. 371–376, Dec. 2020, doi: 10.1109/ISRITI51436.2020.9315377.

[74] E. Marquez-Algaba et al., "COVID-19 Follow-App. Mobile App-Based Monitoring of COVID-19 Patients after Hospital Discharge: A Single-Center, Open-Label, Randomized Clinical Trial," Journal of Personalized Medicine, vol. 12, no. 1, Jan. 2022, doi: 10.3390/JPM12010024.

[75] H. L. Nguyen, K. Tran, P. L. N. Doan, and T. Nguyen, "Demand for Mobile Health in Developing Countries During COVID-19: Vietnamese's Perspectives from Different Age Groups and Health Conditions," Patient Preference and Adherence, vol. 16, pp. 265–284, 2022, doi: 10.2147/PPA.S348790.

[76] M. J. Serrano-Ripoll et al., "Effect of a mobile-based intervention on mental health in frontline healthcare workers against COVID-19: Protocol for a randomized controlled trial," Journal of Advanced Nursing, vol. 77, no. 6, pp. 2898–2907, Jun. 2021, doi: 10.1111/JAN.14813.

[77] T. M. Alanzi et al., "Evaluation of the Mawid mobile healthcare application in delivering services during the COVID-19 pandemic in Saudi Arabia," International Health, vol. 14, no. 2, pp. 142–151, Mar. 2022, doi: 10.1093/INTHEALTH/IHAB018.

[78] L. Mugenyi, R. N. Nsubuga, I. Wanyana, W. Muttamba, N. M. Tumwesigye, and S. H. Nsubuga, "Feasibility of using a mobile App to monitor and report COVID-19 related symptoms and people's

movements in Uganda," PLoS ONE, vol. 16, no. 11 November, Nov. 2021, doi: 10.1371/JOURNAL.PONE.0260269.

[79] P. Echeverría et al., "Monitoring in the workplace for early detection of COVID-19 cases during the COVID-19 pandemic using a mobile health application: COVIDapp.," International Journal of Environmental Research and Public Health, vol. 19, no. 1, Jan. 2022, doi: 10.3390/IJERPH19010167.

[80] G. Feng et al., "PDB28 Innovative Application of a Mobile APP for Gestational Diabetes Health Education in the Era of the COVID-19 Pandemic and BIG DATA," Value in Health, vol. 24, pp. S82–S83, Jun. 2021, doi: 10.1016/J.JVAL.2021.04.426.

[81] J. Kim and B. Gewertz, "Teleurology and digital health app in covid-19 pandemic," Investigative and Clinical Urology, vol. 61, no. 4, pp. 333–334, Jul. 2020, doi: 10.4111/ICU.2020.61.4.333.

[82] J. H. Kim, W. S. Choi, J. Y. Song, Y. K. Yoon, M. J. Kim, and J. W. Sohn, "The role of smart monitoring digital health care system based on smartphone application and personal health record platform for patients diagnosed with coronavirus disease 2019," BMC Infectious Diseases, vol. 21, no. 1, Dec. 2021, doi: 10.1186/S12879-021-05898-Y.

[83] H. Mohammad et al., "Identifying data elements and key features of a mobile-based self-care application for patients with COVID-19 in Iran," Health Informatics Journal, vol. 27, no. 4, pp. 1–15, Dec. 2021, doi: 10.1177/14604582211065703.

[84] X. Wang, C. Markert, and F. Sasangohar, "Investigating Popular Mental Health Mobile Application Downloads and Activity During the COVID-19 Pandemic.," Hum Factors, vol. 00, no. 0, p. 18720821998110, Mar. 2021, doi: 10.1177/0018720821998110.

[85] M. K. Al-Nawayseh, M. AL-Iede, E. Elayeh, R. Hijazeen, K. Al Oweidat, and S. M. Aleidi, "The impact of using a mobile application to improve asthma patients' adherence to medication in Jordan," https://doi.org/10.1177/14604582211042926, vol. 27, no. 3, Sep. 2021, doi: 10.1177/14604582211042926.

[86] A. Otu et al., "Training health workers at scale in Nigeria to fight COVID-19 using the InStrat COVID-19 tutorial app: an e-health interventional study," https://doi.org/10.1177/20499361211040704, vol. 8, Aug. 2021, doi: 10.1177/20499361211040704.

[87] J. C. C. Chow, L. Elizabeth Pathak, and S. T. Yeh, "Using mobile apps in social work behavioral health care service: The case for China," https://doi.org/10.1177/00208728211031953, vol. 64, no. 5, pp. 689–701, Aug. 2021, doi: 10.1177/00208728211031953.

[88] A. Asadzadeh and L. R. Kalankesh, "A scope of mobile health solutions in COVID-19 pandemics," Informatics in Medicine Unlocked, vol. 23, Jan. 2021, doi: 10.1016/J.IMU.2021.100558.

[89] H.-H. Lu, W.-S. Lin, C. Raphael, and M.-J. Wen, "A study investigating user adoptive behavior and the continuance intention to use mobile health applications during the COVID-19 pandemic era: Evidence from the telemedicine applications utilized in Indonesia," Asia Pacific Management Review, Feb. 2022, doi: 10.1016/J.APMRV.2022.02.002.

[90] R. R. Garett, J. Yang, Q. Zhang, and S. D. Young, "An online advertising intervention to increase adherence to stay-at-home-orders during the COVID-19 pandemic: An efficacy trial monitoring individual-level mobility data," International Journal of Applied Earth Observation and Geoinformation, vol. 108, p. 102752, Apr. 2022, doi: 10.1016/J.JAG.2022.102752.

[91] S. Abbaspur-Behbahani, E. Monaghesh, A. Hajizadeh, and S. Fehresti, "Application of mobile health to support the elderly during the COVID-19 outbreak: A systematic review," Health Policy and Technology, vol. 11, no. 1, p. 100595, Mar. 2022, doi: 10.1016/J.HLPT.2022.100595.

[92] J. Wu et al., "Mobile health technology combats COVID-19 in China," Journal of Infection, vol. 82, no. 1, pp. 159–198, Jan. 2021, doi: 10.1016/J.JINF.2020.07.024.

[93] L. S. E. Li, L. L. Wong, and K. Y. L. Yap, "Quality evaluation of stress, anxiety and depression apps for COVID-19," Journal of Affective Disorders Reports, vol. 6, p. 100255, Dec. 2021, doi: 10.1016/J.JADR.2021.100255.

[94] N. S. Alharbi, A. S. AlGhanmi, and M. Fahlevi, "Adoption of Health Mobile Apps during the COVID-19 Lockdown: A Health Belief Model Approach," International Journal of Environmental Research and Public Health, vol. 19, no. 7, p. 4179, Mar. 2022, doi: 10.3390/IJERPH19074179.

[95] A. Bassi, S. Arfin, O. John, and V. Jha, "An overview of mobile applications (apps) to support the coronavirus disease 2019 response in India," Indian Journal of Medical Research, vol. 151, no. 5, pp. 468–473, May 2020, doi: 10.4103/ijmr.IJMR_1200_20.

[96] N. R. Smoll, J. Walker, and G. Khandaker, "The barriers and enablers to downloading the COVIDSafe app – a topic modelling analysis," Australian and New Zealand Journal of Public Health, vol. 45, no. 4, pp. 344–347, Aug. 2021, doi: 10.1111/1753-6405.13119.

[97] A. E. Fischer, T. Van Tonder, Siphamandla B Gumede, and S. T. Lalla-Edward, "Changes in perceptions and use of mobile technology and health communication in south africa during the COVID-19 lockdown: Cross-sectional survey study," JMIR Formative Research, vol. 5, no. 5, May 2021, doi: 10.2196/25273.

[98] R. I. Helou, C. M. Waltmans–den Breejen, J. A. Severin, M. E. J. L. Hulscher, and A. Verbon, "Use of a smartphone app to inform healthcare workers of hospital policy during a pandemic such as COVID-19: A mixed methods observational study," PLoS ONE, vol. 17, no. 1 January, Jan. 2022, doi: 10.1371/JOURNAL.PONE.0262105.

[99] S. Salehinejad, S. R. Niakan Kalhori, S. Hajesmaeel Gohari, K. Bahaadinbeigy, and F. Fatehi, "A review and content analysis of national apps for COVID-19 management using Mobile Application Rating Scale (MARS)," Informatics for Health and Social Care, vol. 46, no. 1, pp. 42–55, 2021, doi: 10.1080/17538157.2020.1837838.

[100] S. Davalbhakta et al., "A Systematic Review of Smartphone Applications Available for Corona Virus Disease 2019 (COVID19) and the Assessment of their Quality Using the Mobile Application Rating Scale (MARS)," Journal of Medical Systems, vol. 44, no. 9, Sep. 2020, doi: 10.1007/S10916-020-01633-3.

[101] B. K. Prahani, B. Jatmiko, B. Hariadi, M. J. D. Sunarto, T. Sagirani, and T. Amelia, "Development Blended Web Mobile Learning Model on COVID-19 Pandemic," TEM Journal, vol. 10, no. 4, pp. 1879–1883, 2021, doi: 10.18421/TEM104-51.

[102] P. S. Peixoto, D. Marcondes, C. Peixoto, and S. M. Oliva, "Modeling future spread of infections via mobile geolocation data and population dynamics. An application to COVID-19 in Brazil," PLoS ONE, vol. 15, no. 7 July, Jul. 2020, doi: 10.1371/JOURNAL.PONE.0235732.

[103] D. D. Satre, M. C. Meacham, L. D. Asarnow, W. S. Fisher, L. R. Fortuna, and E. Iturralde, "Opportunities to Integrate Mobile App–Based Interventions Into Mental Health and Substance Use Disorder Treatment Services in the Wake of COVID-19," American Journal of Health Promotion, vol. 35, no. 8, pp. 1178–1183, Nov. 2021, doi: 10.1177/08901171211055314.

[104] P. Kapoor, P. Sengar, and A. Chowdhry, "Prototype of Customized Mobile Application for Obstructive Sleep Apnea (OSA) Risk Assessment During the COVID-19 Pandemic," https://doi.org/10.1177/0301574220982735, vol. 56, no. 1, pp. 91–95, Jan. 2021, doi: 10.1177/0301574220982735.

[105] L. Cosgrove, J. M. Karter, Z. Morrill, and M. McGinley, "Psychology and Surveillance Capitalism: The Risk of Pushing Mental Health Apps During the COVID-19 Pandemic," https://doi.org/10.1177/0022167820937498, vol. 60, no. 5, pp. 611–625, Jun. 2020, doi: 10.1177/0022167820937498.

[106] T.-C. T. Chen and C.-W. Lin, "An FGM decomposition-based fuzzy MCDM method for selecting smart technology applications to support mobile health care during and after the COVID-19 pandemic," Applied Soft Computing, vol. 121, p. 108758, May 2022, doi: 10.1016/J.ASOC.2022.108758.

[107] A. Verma, S. B. Amin, M. Naeem, and M. Saha, "Detecting COVID-19 from chest computed tomography scans using AI-driven android application," Computers in Biology and Medicine, vol. 143, p. 105298, Apr. 2022, doi: 10.1016/J.COMPBIOMED.2022.105298.

[108] C. F. Munive-Aponte, J. J. Dávila-Asto, and G. Tirado-Mendoza, "Design of an M-Commerce mobile application to reduce the cessation of operations of textile companies due to the social isolation generated by SARS-CoV-2 in Peru," Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology, vol. 2021-July, 2021, doi: 10.18687/LACCEI2021.1.1.115.

[109] Y. A. Daineko, D. D. Tsoy, A. M. Seitnur, and M. T. Ipalakova, "Development of a Mobile e-Learning Platform on Physics Using Augmented Reality Technology," International Journal of Interactive Mobile Technologies, vol. 16, no. 5, pp. 4–18, 2022, doi: 10.3991/IJIM.V16I05.26961.

[110] C. Vladoescu, M. Tunea, and L. Stanciu, "Benefits of Using Mobile Applications to Keep Under Control, Detect, Attenuate and Monitor COVID-19 Pandemic," pp. 1–4, Dec. 2020, doi: 10.1109/EHB50910.2020.9280229.

[111] D. Chen, A. Bucchiarone, and Z. Lv, "MeetDurian: A Gameful Mobile App to Prevent COVID-19 Infection," Proceedings - 2021 IEEE/ACM 8th International Conference on Mobile Software Engineering and Systems, MobileSoft 2021, pp. 69–72, May 2021, doi: 10.1109/MOBILESOFT52590.2021.00016.

[112] A. Althunibat, F. Altarawneh, R. Dawood, and M. A. Almaiah, "Propose a New Quality Model for M-Learning Application in Light of COVID-19," Mobile Information Systems, vol. 2022, 2022, doi: 10.1155/2022/3174692.

[113] Y. Fan, Z. Wang, S. Deng, H. Lv, and F. Wang, "The function and quality of individual epidemic prevention and control apps during the COVID-19 pandemic: A systematic review of Chinese apps," International Journal of Medical Informatics, vol. 160, Apr. 2022, doi: 10.1016/J.IJMEDINF.2022.104694.

[114] P. Mehta, S. L. Moore, S. Bull, and B. M. Kwan, "Building MedVenture - A mobile health application to improve adolescent medication adherence - Using a multidisciplinary approach and academic-industry collaboration.," Digit Health, vol. 7, p. 20552076211019876, May 2021, doi: 10.1177/20552076211019877.

[115] J. Burrieza-Galán et al., "A methodology for understanding passenger flows combining mobile phone records and airport surveys: Application to Madrid-Barajas Airport after the COVID-19 outbreak," Journal of Air Transport Management, vol. 100, p. 102163, May 2022, doi: 10.1016/J.JAIRTRAMAN.2021.102163.

[116] N. S. Alharbi, N. Alsubki, S. R. Altamimi, W. Alonazi, and M. Fahlevi, "COVID-19 Mobile Apps in Saudi Arabia: Systematic Identification, Evaluation, and Features Assessment," Frontiers in Public Health, vol. 10, Mar. 2022, doi: 10.3389/fpubh.2022.803677.

[117] P. Tangtisanon, "COVID-19 Pandemic Prevention Mobile Application for on Campus Classroom," 2021 IEEE 6th International Conference on Computer and Communication Systems, ICCCS 2021, pp. 1117–1121, Apr. 2021, doi: 10.1109/ICCCS52626.2021.9449201.

[118] I. O. Yahuarcani et al., "Mobile application for the dissemination, learning and revitalisation of the native Muniche language in the Peruvian Amazon, in the context of Covid-19," Proceedings - 2021 4th International Conference on Inclusive Technology and Education, CONTIE 2021, pp. 82–88, 2021, doi: 10.1109/CONTIE54684.2021.00023.

[119] M. A. Amrein, G. G. Ruschetti, C. Baeder, M. Bamert, and J. Inauen, "Mobile intervention to promote correct hand hygiene at key times to prevent COVID-19 in the Swiss adult general population: study protocol of a multiphase optimisation strategy," BMJ Open, vol. 12, no. 3, p. e055971, Mar. 2022, doi: 10.1136/BMJOPEN-2021-055971.

[120] S. S. Kaware, M. K. Gupta, and A. K. Gupta, "Mobile Phone Educational Applications: Their Importance in Academic Learning During Covid 19 Pandemic," International Journal of Early Childhood Special Education, vol. 13, no. 2, pp. 1013–1020, 2021, doi: 10.9756/INT-JECSE/V13I2.211144.

[121] J. Hodges et al., "Six-month outcomes of the HOPE smartphone application designed to support treatment with medications for opioid use disorder and piloted during an early statewide COVID-19 lockdown," Addiction Science and Clinical Practice, vol. 17, no. 1, Dec. 2022, doi: 10.1186/s13722-022-00296-4.

[122] K. Kostyrka-Allchorne et al., "Supporting Parents & Kids Through Lockdown Experiences (SPARKLE): A digital parenting support app implemented in an ongoing general population cohort study during the COVID-19 pandemic: A structured summary of a study protocol for a randomised controlled trial," Trials, vol. 22, no. 1, Dec. 2021, doi: 10.1186/s13063-021-05226-4.

[123] M. Whaiduzzaman et al., "A Privacy-Preserving Mobile and Fog Computing Framework to Trace and Prevent COVID-19 Community Transmission," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 12, pp. 3564–3575, Dec. 2020, doi: 10.1109/JBHI.2020.3026060.

[124] S. Hisada et al., "Surveillance of early stage COVID-19 clusters using search query logs and mobile device-based location information," Scientific Reports, vol. 10, no. 1, Dec. 2020, doi: 10.1038/S41598-020-75771-6.

[125] J. Baik and E. Jang, "Where horizontal and vertical surveillances meet: Sense-making of US COVID-19 contact-tracing apps during a health crisis," https://doi.org/10.1177/20501579221078674, Feb. 2022, doi: 10.1177/20501579221078674.

[126] F. Kurtaliqi, M. Zaman, and R. Sohier, "The psychological reassurance effect of mobile tracing apps in Covid-19 Era," Computers in Human Behavior, vol. 131, Jun. 2022, doi: 10.1016/j.chb.2022.107210.

[127] R. Dhull, D. Chava, D. V. Kumar, K. M. V. V. Prasad, G. Samudrala, and M. V. Bhargav, "Pandemic Stabilizer using Smartwatch," 2020 International Conference on Decision Aid Sciences and Application, DASA 2020, pp. 860–866, Nov. 2020, doi: 10.1109/DASA51403.2020.9317056.

[128] A. Asadzadeh, T. Samad-Soltani, and P. Rezaei-Hachesu, "Applications of virtual and augmented reality in infectious disease epidemics with a focus on the COVID-19 outbreak," Informatics in Medicine Unlocked, vol. 24, p. 100579, Jan. 2021, doi: 10.1016/J.IMU.2021.100579.

[129] S. S. Vedaei et al., "COVID-SAFE: An IoT-based system for automated health monitoring and surveillance in post-pandemic life," IEEE Access, vol. 8, pp. 188538–188551, 2020, doi: 10.1109/ACCESS.2020.3030194.

[130] H.-H. Lu, W.-S. Lin, C. Raphael, and M.-J. Wen, "A study investigating user adoptive behavior and the continuance intention to use mobile health applications during the COVID-19 pandemic era: Evidence from the telemedicine applications utilized in Indonesia," Asia Pacific Management Review, Feb. 2022, doi: 10.1016/J.APMRV.2022.02.002.

[131] S. Abbaspur-Behbahani, E. Monaghesh, A. Hajizadeh, and S. Fehresti, "Application of mobile health to support the elderly during the COVID-19 outbreak: A systematic review," Health Policy and Technology, vol. 11, no. 1, Mar. 2022, doi: 10.1016/J.HLPT.2022.100595.

[132] G. L. Vasconcelos, G. C. Duarte-Filho, A. A. Brum, R. Ospina, F. A. G. Almeida, and A. M. S. Macêdo, "Situation of COVID-19 in Brazil in August 2020: An Analysis via Growth Models as Implemented in the ModInterv System for Monitoring the Pandemic," Journal of Control, Automation and Electrical Systems, vol. 33, no. 2, pp. 645–663, Apr. 2022, doi: 10.1007/s40313-021-00853-3.

[133] D. D. Satre, M. C. Meacham, L. D. Asarnow, W. S. Fisher, L. R. Fortuna, and E. Iturralde, "Opportunities to Integrate Mobile App–Based Interventions Into Mental Health and Substance Use Disorder Treatment Services in the Wake of COVID-19," American Journal of Health Promotion, vol. 35, no. 8, pp. 1178–1183, Nov. 2021, doi: 10.1177/08901171211055314.

[134] J. Ispahany and R. Islam, "Detecting malicious COVID-19 URLs using machine learning techniques," 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events, PerCom Workshops 2021, pp. 718–723, Mar. 2021, doi: 10.1109/PERCOMWORKSHOPS51409.2021.9431064.

[135] H. C. Sun, X. F. Liu, Z. W. Du, X. K. Xu, and Y. Wu, "Mitigating COVID-19 Transmission in Schools with Digital Contact Tracing," IEEE Transactions on Computational Social Systems, vol. 8, no. 6, pp. 1302–1310, Dec. 2021, doi: 10.1109/TCSS.2021.3073109.

[136] R. I. Helou, C. M. Waltmans–den Breejen, J. A. Severin, M. E. J. L. Hulscher, and A. Verbon, "Use of a smartphone app to inform healthcare workers of hospital policy during a pandemic such as COVID-19: A mixed methods observational study," PLoS ONE, vol. 17, no. 1 January, Jan. 2022, doi: 10.1371/JOURNAL.PONE.0262105.

[137] S. Abbaspur-Behbahani, E. Monaghesh, A. Hajizadeh, and S. Fehresti, "Application of mobile health to support the elderly during the COVID-19 outbreak: A systematic review," Health Policy and Technology, vol. 11, no. 1, p. 100595, Mar. 2022, doi: 10.1016/J.HLPT.2022.100595.

[138] O. Tutsoy, "COVID-19 Epidemic and Opening of the Schools: Artificial Intelligence-Based Long-Term Adaptive Policy Making to Control the Pandemic Diseases," IEEE Access, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3078080.

[139] E. Jordan, D. E. Shin, S. Leekha, and S. Azarm, "Optimization in the Context of COVID-19 Prediction and Control: A Literature Review," IEEE Access, vol. 9, pp. 130072–130093, 2021, doi: 10.1109/ACCESS.2021.3113812.

[140] H. J. L. Singh, D. Couch, and K. Yap, "Mobile Health Apps That Help With COVID-19 Management: Scoping Review," JMIR Nursing, vol. 3, no. 1, p. e20596, Aug. 2020, doi: 10.2196/20596.

[141] A. A. Khalil, Meyliana, A. N. Hidayanto, and H. Prabowo, "Identification of Factor Affecting Continuance Usage Intention of mHealth Application : A Systematic Literature Review," ICICoS 2020 - Proceeding: 4th International Conference on Informatics and Computational Sciences, Nov. 2020, doi: 10.1109/ICICOS51170.2020.9299038.

[142] H. Kondylakis et al., "COVID-19 Mobile Apps: A Systematic Review of the Literature," J Med Internet Res, vol. 22, no. 12, Dec. 2020, doi: 10.2196/23170.

[143] A. Alshawi et al., "Data privacy during pandemics: a systematic literature review of COVID-19 smartphone applications," PeerJ Computer Science, vol. 7, p. e826, Jan. 2021, doi: 10.7717/PEERJ-CS.826/FIG-7.

Modelling of IoT-WSN Enabled ECG Monitoring System for Patient Queue Updation

Parminder Kaur[1], Hardeep Singh Saini[2], Bikrampal Kaur[3]
Research Scholar, Electronics Engineering Department, IKGPTU, Jalandhar, Punjab, India[1]
Prof. ECED, Indo Global College of Engineering Punjab, India[2]
Prof. CSED, CEC Landran, Punjab, India[3]

***Abstract*—The advancement of communication technologies has led to the interconnection of different sensors using the Internet of Things (IoT) and Wireless Sensor Network (WSN). WSN for healthcare applications has expanded exponentially due to evolving advantages such as low power requirement of sensors, transmission accuracy, and cost-efficiency. For heart attack patients, the future lies in ECG monitoring in which wearable sensors can be used to acquire patient information. In this paper, an attempt has been made to develop a novel IoT-enabled WSN to record patient information for detection of heart attack and to update queue of patients to ensure prioritized medical attention to critical patients. In the WSN, the Rayleigh Fading channel has been used to transmit data that can be accessed using the cloud repository by the medical staff remotely. The distance from the patient to the medical staff is calculated using Euclidean distance. Further, SNR in comparison to throughput and BER has been computed. The higher SNR indicates the maximum information transfer from patient to hospital staff. The proposed system uses the Grasshopper Optimization and CBNN based disease classification system and bubble sort algorithm has been used for updating patient queue. The proposed GHOA and CBNN has shown improved accuracy of 2.14% over existing techniques like CNN which has accuracy around 82% for R-R feature selection of ECG signals as compared to 82.72% shown by GHOA-CBNN.**

***Keywords*—WSN; cloud; ECG monitoring; wearable sensors; IoT; queue updation**

I. Introduction

Wireless Sensor Network (WSN) consists of various sensors used to sense the information and process the same for different applications. Internet of Things (IoT) can allow a seamless communication between patient and medical staff, by transferring data from a WSN to cloud computing platforms to always ensure uninterrupted health monitoring. In healthcare applications [1], availability of wearable sensors allows the continuous monitoring of patient parameters, with distinctive threshold levels for serious conditions such as heart failure, pulse rate, diabetes, and many others. Wearable devices enable long-term, continual assessment of the patient's critical indicators while allowing them total freedom of movement. The WSN-based healthcare systems use biosensors to collect physiological information from patients. The collected information can be shared using the Wireless networks directly with the server or doctors for clinical review. A promising technology in the healthcare field, Wireless Body Area Network (WBAN) offers higher-quality applications and services. Consequently, a more trustworthy analysis may be carried out by the doctors [2] using this vast amount of data rather than relying on the one recorded during a brief stay in the hospital. WBAN is a sophisticated monitoring system made from computing-capable wearable and implantable nodes that are positioned in, on, and nearby a person's body. This fastens the decision-making process and is useful for quick monitoring of the patient. Moreover, the integration of IoT with WBAN-enabled applications significantly reduce the cost of travel, and time, especially for long-term applications of monitoring in which doctors wait for a long time to record the ECG patterns. Researchers in [3] have been actively engaged in using the IoT and wearable sensor technology for the detection of cardiovascular diseases.

IoT-WSN based ECG monitoring systems in [4] are one of the powerful technological advancements to monitor the health of the patient remotely in real-time. An IoT-based ECG monitoring system enables the collection and analysis of ECG data for remotely monitoring patients. The data collected wirelessly can be directly stored or processed using the cloud computing devices. To create a decision support system that could aid in early diagnosis and treatment, the medical staff in [5] uses this data for additional analysis which can result in saving precious lives.

Heart disease is the leading cause of mortality worldwide. Hence, there is a need for the development of intelligent tools to detect heart-related diseases timely and accurately using a low-cost device. An IoT-WSN enabled ECG monitoring system has been developed in [6] which is a widely accepted method for the diagnosis of heart-related diseases. Conventional 12- electrode ECG monitoring systems are bulky and non-portable making it mandatory for the patient to be in the hospital for the process. A survey conducted in [7] presented trends and techniques related to IoT in healthcare applications. Researchers in [8] integrate the ECG monitoring and classification using IoT and deep neural networks that make the process easier and faster.

In contrast to the conventional approach presented in [9], 3 or 5-electrode ECG devices are now often utilized because they can provide precise ECG signals. These IoT-based heart rate monitoring sensors are portable that collect patient ECG signals and send the information to a mobile application via a wireless communication module. Numerous monitoring and analytic tools for heart rate monitoring such as RR wave peaks analysis were described in literature and related devices are being introduced in [10] and [11] for implementation. Additionally, several methods had been developed for peak

detection, such as the Hidden Markov model and Pan Tompkin's. The major issues with such methods are absence of standardized features, lack of robustness, real-time monitoring of ECG samples, and portability, and lack of sustainable solutions and there is a need for medical acceptance for the analysis of the signal.

Therefore, understanding these limitations, there is a need of WSN enabled ECG monitoring device that is faster and medically accepted. The proposed framework in this paper deals with acquisition of ECG data of a patient using wearable sensors which is transmitted to the medical team using the Rayleigh Fading channel, data can be stored and processed using cloud devices where a R-R peak analysis is performed. The R-R peak interval in healthy individuals ranges from 0.6 to 1.2 seconds. Any variation in the R-R interval helps to identify heart disease conditions. The proposed model uses the MIT-BIH Arrhythmia Dataset using five different classes such as N, S, V, F and Q. The main contribution of this paper is as follows:-

- Modelling of IoT-WSN enabled ECG monitoring system that works for remote locations.
- The ECG signal acquired through wearable sensors is transmitted from the patient to the allotted medical advisor using the Rayleigh Fading Channel and evaluation of parameters like throughput, Signal to Noise Ratio (SNR) and Bit Error Rate (BER) has been done.
- An Application is developed for remote as well as local supervision for patients in which proposed system displays the initial queue and updated queue of patients based on their seriousness levels.

The main motivation of this article is to present an ECG monitoring system using the integrated IoT-WSN technology for a healthcare monitoring system that can address the challenges of the existing systems. The advancement in wearable sensor technology allows the practitioners to use the WSN and IoT for the development of cost effective and reliable patient monitoring system. The present study is a continuation of research conducted in [12] to collect the vital signs using the wearable sensors. Furthermore, the real motivation behind this research is to save time, patients have to wait for long hours in hospitals for medical care and in critical cases that can prove fatal also. So, with the development of this WSN-IoT enabled ECG monitoring application, queue is updated based on the patient seriousness, and thus ensuring immediate and precise medical care to the patient.

The organization of the article is as follows: Section II details the related work to discuss the existing techniques for ECG monitoring. The next section discusses the research methodology in which different techniques adopted for communication has been detailed. Results and discussion are illustrated in section IV and the conclusion is given in section V.

II. Related Work

The authors in [13] proposed an IoT-based ECG monitoring for health care applications. The authors used the ECG sensor and development boards to send the information to remote locations. The Bluemix device had been used in conjunction with MQTT (Message Queuing Telemetry Transport) for the integration of different types of devices. The use of this protocol supports Machine to Machine communication without human intervention. The main advantage of the proposed system is its low cost but it also has limited ability to provide the results in a controlled manner.

The research conducted by Satija et al. [14] in which a novel ECG telemetry system had been developed for continuous cardiac health monitoring applications that is IoT enabled and lays emphasis on signal quality. The implementation of this work had been done using the ECG sensors, Arduino, Android phone, Bluetooth, and a cloud server. The interconnection of these devices enables the authors to create and build a lightweight ECG monitoring device for automatically categorizing the acquired ECG signal into acceptable or unacceptable classes and to implement Sensor enabled ECG monitoring program in real-time.

The study conducted by Gogate and Bakal [15] used the WSN for the development of the three-tier architecture for a healthcare monitoring system. The patient parameters such as heart rate, oxygen saturation, and temperature had been measured. The biosensors had been directly connected to the Arduino board to send the information to the server using wireless channel. The emergency patients were notified using alert systems from mobile phones and the accuracy of the developed system was about 95% with a minimum response time of around 10 seconds.

The authors of [16] created a wearable medical device that uses a three-lead ECG sensor to collect ECG data to detect arrhythmias in real-time. To detect arrhythmia and do real-time heart monitoring, this study offers a workable and simple method. It carries out ECG signal interpretation and wirelessly notifies the patient's doctor of arrhythmia at once. For instance, the Pan-Tompkins and adaptive filtering framework were used to find premature ventricular contractions (PVCs), a prevalent kind of arrhythmia. MIT-BIH arrhythmia database benchmark records were used to successfully test the robustness of the research work. The device is low-cost and uses the Raspberry Pi module for communication.

Practitioners in [17] proposed a novel system for the monitoring of remote healthcare using Machine Learning and IoT enabled devices. The authors allow the monitoring in real-time and associate the data with cloud computing. The paper also throws light to evaluate the prediction system for the measurement of heart diseases. The experimental results have been compared using machine learning classifiers such as Decision Tree, Random Forest, Support Vector Machine, and K-nearest neighbor. The highest accuracy of about 57.37% was obtained using Linear-Support Vector Machine. The limitation of the paper is that it is unable to provide the required security level for patient data.

Researchers in [18] integrated the concept of Big Data, IoT, and Nano-electronics to resolve the issue of inconspicuous monitoring. The use of Nano-electronic devices allows the users to send the data to numerous users such as physicians, medical advisors, and caretakers to analyse the data. The

transmission of signal had been done using the sensor considering the communication protocols such as Zigbee and LAN etc. The physicians at the remote location access the data and can view the reports using the sensor devices. The integration of these three technologies allows doctors to fasten the data analysis and decision-making process. The main limitation of this paper is increase in system complexity due to the use of different computational devices and system is less reliable due to the use of Nano tubes.

Huda et al. [19] developed a low-cost and low-power ECG monitoring system in conjunction with a deep learning model to facilitate the automatic detection of arrhythmia cardiovascular disease. The authors used the AD8232 chip to process the ECG signal and Convolutional Neural Network had been used for the classification of MIT-BIH arrhythmia disease. The accuracy of the developed system was 94.03% and provided effective results.

A low-cost ECG monitoring system to measure the seriousness of the patient was developed in [20]. The use of low-energy devices efficiently measures the arrhythmia detection, saturation level of oxygen, and temperature of the body that can be directly sent to the medical advisor via sensor devices. Further, the GSM module had been used to send an alert to the doctors in case of emergency conditions, and a web application equipped with deep learning facilitates the communication process between doctor and patient. The proposed device serves remote areas and is also helpful for telehealth care.

The authors in [21] offered a cloud-based method for remotely monitoring heart disease. To enable data visualization, fast reaction, and long-term connectivity between equipment and users, Hyper Text Transfer Protocol (HTTP) and Message Queuing Telemetry Transport (MQTT) servers had been employed. A communication technology called Bluetooth which relies on low energy (BLE 4.0) had been used to transmit information between a device and a wireless gateway. Filtration methods were used in the developed framework to suppress interruptions, background noise, and motion artifacts. It provides ECG signal analysis to identify several parameters, including beats, QRS complex intervals, PQRST wave, and breathing rate. The designed model was examined and found to be trustworthy for remote ECG monitoring. The main drawback is an inefficient system which is not portable. The research conducted by Ghafil et al. [22] had used medical sensors to collect physiological information from the patients. Wearable sensors had been used for the continuous monitoring of the patient. Holter machine had been used to access the ECG signals and WSN had been used for the transmission of the signal. Moreover, a cloud server had been used to store the recorded signals and the decision-making process finally had been done using medical sensors.

The authors in [25] used the cooperative Nano network for communication using the vivo technology. The authors used the WSN for communication and results show improvement from existing techniques.

The development of ECG monitoring systems has been an extensive field of research for the last many years. The need of timely medical care to cardiovascular patients motivated the authors to present a mechanism in the present paper that can overcome the delay in getting medical care. The study of literature and analysis of numerous projects related to this subject helped authors conclude that the queue Updation system has not yet been developed. This paper thus presents a novel system using IoT-WSN enabled ECG monitoring with the use of cloud repository, to maintain a queue based on the seriousness of the patient.

III. Research Methodology

The system architecture is divided into two phases in which phase 1 includes the use of WSN architecture to sense and collect the information. The phase 2 includes the storage and processing of information on the cloud architecture.

In Phase 1, the interconnection of various devices such as sensors, cloud processors, etc. using IoT for the monitoring of ECG signal for the detection of arrhythmia diseases is shown in Fig. 1. WSN mainly consists of wearable sensors employed for the detection of ECG signal and the Wireless communication channel such as the Rayleigh fading channel that has been used to transfer the information from the remote location to the hospital. The real time implementation of this work will involve placing ECG sensors on the patient's body. The different types of wearable sensors used to acquire the information are already discussed in article [12].

The data set used for this research is the MIT BIH Arrythmia dataset, which is accessed from the Kaggle link mentioned below: (https://www.kaggle.com/datasets/shayanfazeli/heartbeat).

MIT-BIH arrhythmia database, The MIT BIH dataset consists of ECG recordings of 47 different subjects recorded at a sampling rate of 360 Hz. The MIT-BIH dataset includes five different types of classes N, S, V, F and Q which are labeled as 0, 1, 2, 3, 4 for current study. The main aim is to select the relevant attributes and then train the data using the appropriate training algorithm. Grasshopper Optimization algorithm and CBNN have been used to determine the relevant attributes from the Kaggle ECG dataset [24].

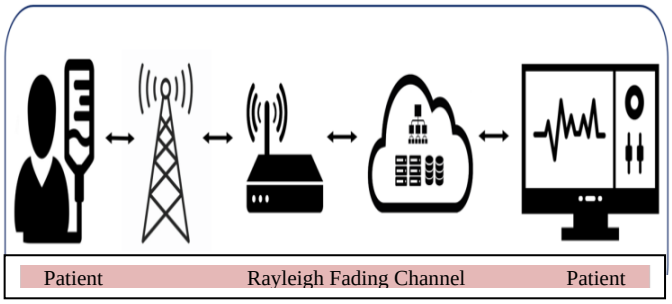


Fig. 1. IoT WSN Patient Monitoring System

A. WSN Framework

The proposed WSN framework includes the registration of patients, active patients, and initialization of network, determining the patient information, initiating the medical procedure, and associating medical staff with active patients considering distance between the two.

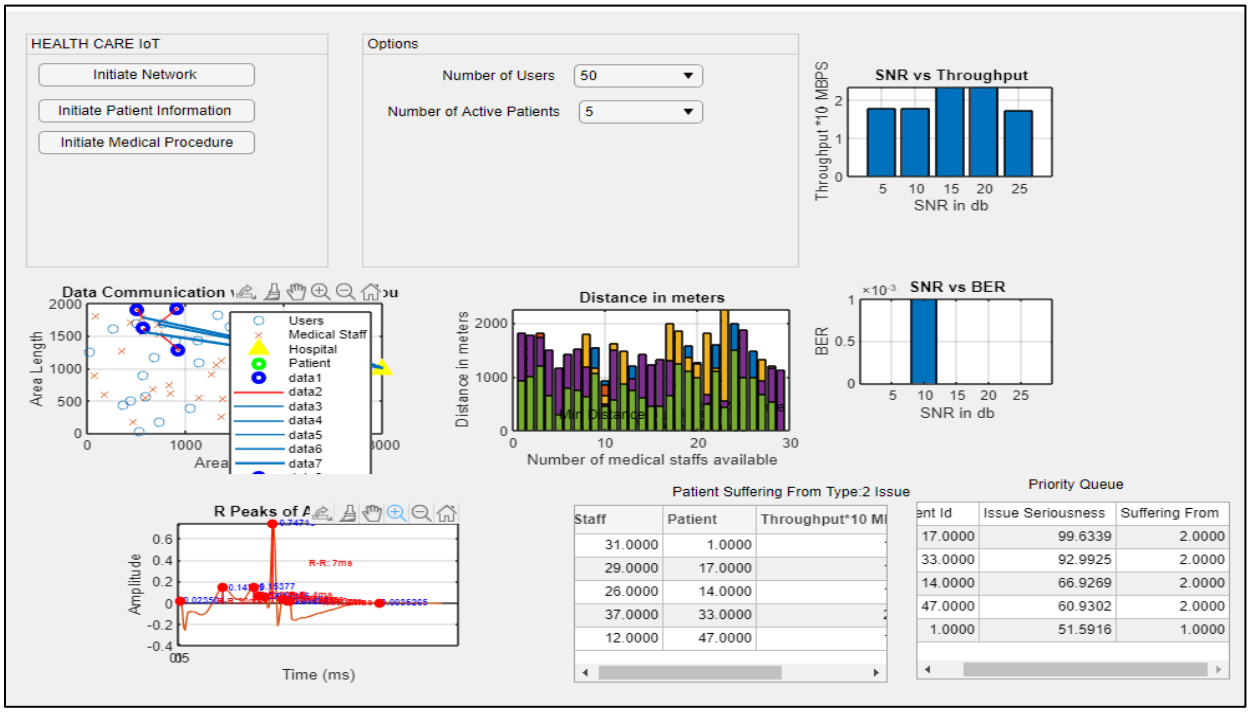


Fig. 2. The Proposed Framework IoT-WSN Enabled Healthcare System.

The proposed framework is shown in Fig. 2. In the developed framework, it is seen that number patients who have been registered is 50 and the queue is updated by experimenting with different or same number of active patients during different iterations of the program. When the initiate network button is pressed then information about the patients is recorded. The entire information is stored and processed using the cloud technology. When the initiate medical procedure button is pressed, then the RR peak analysis and sorting of patients based on seriousness starts and updated queue is displayed along with the initial queue in which staff allocation to patients and throughput and BER of the channel are displayed for each patient.

Another important information that is displayed is the classification of heart disease based on RR peak analysis, patients are classified as being suffering from disease type 0,1, 2, 3 and 4 that corresponds to the N, S, V, F and Q categories of MIT BIH Arrhythmia database respectively.

B. Rayleigh Fading Channel

Rayleigh fading is a model for describing the kind of fading that happens when multipath propagation is present. The Rayleigh fading channel has been used to transmit the ECG information of patient in the form of RR peaks from a remote location. However, the studies conducted in [23] used the Rayleigh Fading channel with Internet of Medical Technology for the transmission of signals. The integrated technology helps the practitioners to transmit the patient information for better communication. This channel is used for amplification and transmitting or relaying the signal, a simple technique that can be employed with a lesser number of associated overhead bits and therefore chosen for implementation and analysis of the proposed work. The parameters of Rayleigh channel are shown in Table 1.

TABLE I. PARAMETERS OF RAYLEIGH CHANNEL

Parameter Name	Value
Input symbol rate	9600
Number of samples per input symbol	10
Input sampling frequency (samples/s)	9600*10
Input sampling period (s)	1/Fs
Number of input symbols to simulate	1e6
Number of channel samples to simulate	1e6*10
Maximum Doppler frequency shift (Hz)	100
Number of samples of auto-covariance	5000
Number of transmit antennas	1
Number of receive antennas	1
Number of sinusoids	48
Frame length	10000
Number of frames	1e6*10/10000

C. Channel Parameters

Specifically, there are two nodes - Source node and the Destination node. The simple computation and detection process has been done using the source node and data is transmitted using the relay nodes. The relay node receives the signal, amplifies the signal, and then simply forwards the signal further to the destination node which is located at a certain distance. The destination node transmits the information through the Rayleigh Fading channel to a remote hospital for the monitoring of the signal. The received data from the channel is analyzed using the Bit Error Rate (BER) and Throughput of the signal which is also referred to as signal strength. The mathematical representation for throughput and BER is as follows.

$$Throughput = \frac{t_p}{pl+I} \quad (1)$$

Where t_p is the transmission power, pl is the path loss, and 'I' is the interference in the system as in (1). To calculate path loss, the following mathematical representation is used:

$$pl = 32.4 + 21 * \log(d_{2_h}) + 20 * \log10(fc) \quad (2)$$

Where d_{2_h} is the distance to the hospital from the user, fc is the central frequency viz. 3.5 GHz as in (2). BER is simply calculated by looking up a total false bit received to total sent bits.

Considering, that there are 'm' number of serious patients and 'n' number of relay nodes make the cooperative network for communication, the net SNR computed at the destination end has been determined as in equation 3[25].

$$SNR_{net} = \sum_{i=1}^{m} \frac{SNR_{source} SNR_{Destination}}{SNR_{source} + SNR_{Destination} + 1} \quad (3)$$

SNR has been computed using equation 3

$$SNR(dB) = 10 \times \log\left(\frac{Power_{transmission}(T_P) - Path\ Loss\ (P_L)}{Power_{noise}(P_N)}\right) \quad (4)$$

T_P is recommended to be less than 1mW in this case and Path Loss has been computed using equation 2.

$$f = \frac{c}{\sqrt{\epsilon} \times \lambda} \quad (5)$$

ϵ is the permittivity and it is about 0.2625 for human tissue near the surface of the skin [2] and further, the comparative study of SNR has been done with throughput. The lower value of path loss and higher value of SNR signifies an error-free transmission process at the destination end. The collected information is stored in the cloud technology that can be further accessed by the doctors and medical advocates.

D. ECG Data Analysis and Queue Updation

After the computation of SNR in relation to throughput, the classification process has been carried out. The medical practitioners have been allotted based on the seriousness of the patient. The data communication with a hospital as shown in Fig. 2 has been taken place and the distance between the patient and medical staff has been computed.

The distance from the patient to the medical staff is calculated using Euclidean distance which is defined as follows.

$$d = \sqrt{[(ux - mx)^2 + (uy - my)^2]} \quad (6)$$

where ux, uy, mx, and my are the geostationary coordinates of the ECG Data.

These are related to the patient and the medical staff. The nearest medical representative to the patient has been allotted accordingly. The patient is treated by extracting the RR features; the disease is classified using the Grasshopper Optimization algorithm (GHOA). The optimization algorithm plays a significant role in selecting the features having maximum information. The algorithm works on the behavior of grasshopper for both exploration and exploitation phase, and use the novel fitness function used to select the RR features. Further, training and testing has been done in which 70% data is trained using Conjugate based Neural Network (CBNN) and same technique is used for classification of 30% test data. The proposed technique was compared against CNN; GHOA-CBNN resulted in 82.72% accuracy and CNN was at 80.58%, resulting in an improvement for R-R feature selection accuracy by 2.14%.

When the patient data is received at the receiving end, it is first processed for the R-R peak analysis and then furthermore the type of issue is determined that a patient is suffering from. Here in this scenario in Fig. 3, the classification engine compares the supplied R-R peak values against the stored values. The matching score is calculated based on the correlation of the supplied data to the repository data. A high matching score represents more seriousness in the patient data whereas a low matching score represents less seriousness in the patient data. The patient with a high matching score is treated before any other patient if there is no other patient that has a higher matching score than that of the current patient. To update the queue, the proposed algorithm utilizes the factor of seriousness for the patient using the following equation (7).

$$Seriousness\ of\ Patient = (Categorized\ class\ score - Original\ Score) * 100 \quad (7)$$

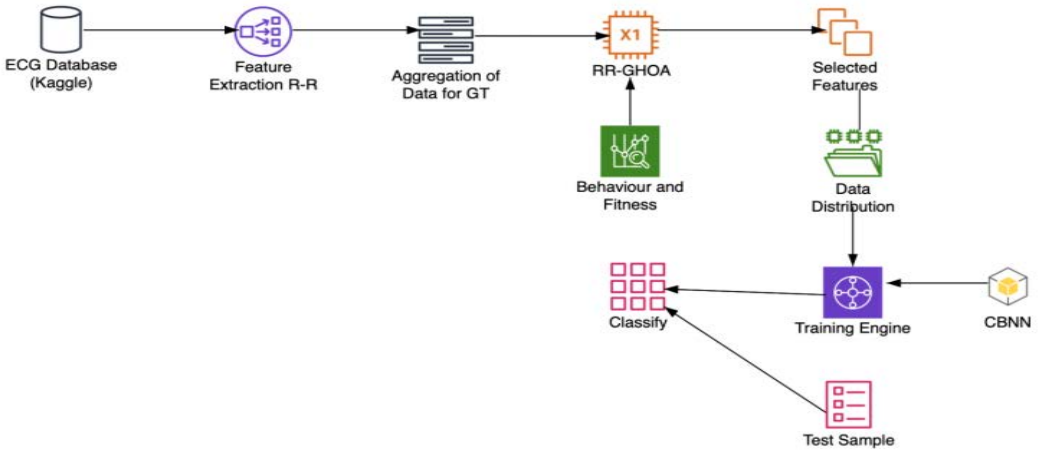


Fig. 3. Analysis and Disease Classification

When the classified result is attained, the disease is decided on the basis of the classification score that represents the matching value with the original class.

If the classified score is close to the original score, the patient is close to the exact disease and hence requires special attention at the very same time.

IV. Results and Discussion

The proposed work is analyzed using wearable sensors to monitor the ECG signal. The Rayleigh fading channel is used for wireless communication from the source node to the destination node. The information is transmitted with the preamble of 32 bits and a payload of 80 bytes of information transmitted using the ECG data acquisition system. The simulation has been performed for a different number of patients registered for monitoring of ECG. The simulation has been performed with SNR ranging from – 10 to 30 dB. Further, throughput and Bit Error Rate (BER) analysis have been determined considering the proposed framework. In this paper, the queue is updated based on the seriousness of the patients. The seriousness is organized in descending order and patients are treated in that manner only. Staff allocation to patients and throughput and BER is listed in Table II. The serious patients have been prioritized and arranged accordingly in the list and then the queue is updated as shown in Table III.

Table II shows the throughput for different patients and BER is listed which is 0.0002 for almost all patients. The throughput value varies from patient to patient. For instance, patient id 1 has throughput of 1.98 Mbps and for 47th patient id, it is 1.99 Mbps as shown in Fig. 4.

Table III shows the arrangement of patient queue based on the level of seriousness and different class of disease. It is seen that patient id 17 having maximum level of seriousness and thus arranged on the top in the queue. Patient id 14 suffered from type 2 disease had a seriousness of 66.92%, is listed as number 3 in the list. Patient id 1 suffered from type 1 disease and had a seriousness of 51.5916%, is queued at number 5.

TABLE II. Medical Staff Allocation to Patients

Staff ID	Patient ID	Throughput*Mbps	BER
31	1	1.9893	0.0002
29	17	1.9947	0.0002
26	14	1.9860	0.0002
37	33	2.0560	0.0002
12	47	1.9971	0.0002

TABLE III. Arranging the Patient's Queue based on Seriousness

Patient ID	Seriousness	Class of Disease
17	99.6339	2
33	92.9925	2
14	66.9269	2
47	60.9302	2
1	51.5916	1

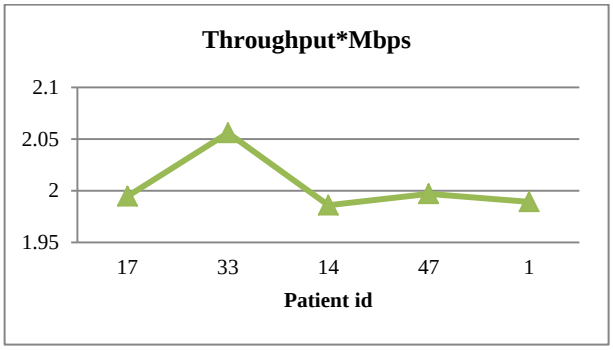


Fig. 4. Throughput of Different Patients.

V. Conclusion

In this paper, we proposed a model for healthcare applications which is based on IOT-WSN architecture. The ECG signal can be collected using the wireless wearable sensors but for this paper MIT BIH Arrhythmia dataset has been considered. The IoT-enabled framework performs better in terms of throughput and BER. The use of the optimization approach classifies the disease efficiently and updates patient queue based on the seriousness level of the patient. The presented study is simple and convenient and its real time implementation can help save precious lives. The proposed IoT-WSN system uses less power, and the complete framework has been simulated in MATLAB. This work can be extended further to display the queue in the hospitals that are visible not only to doctors but also to the patients and staff and such a system can be developed for disease specific applications or for random patients reporting in the emergency OPD.

References

[1] P. P. Ray, D. Dash, and D. De, “Internet of things-based real-time model study on e-healthcare: Device, message service and dew computing,” Comput. Networks, vol. 149, pp. 226–239, 2019, doi: 10.1016/j.comnet.2018.12.006.

[2] N. Arora, S. H. Gupta, and B. Kumar, “Analyzing and Optimizing Cooperative Communication for in Vivo WBAN,” Wirel. Pers. Commun., vol. 122, no. 1, pp. 429–450, 2022, doi: 10.1007/s11277-021-08906-1.

[3] Zhadyra N. Alimbayeva et al, “Portable ECG Monitoring System,” Intl. J. of Adv. Computer Science and Applications., vol. 13, no. 4, pp. 64-76, 2022.

[4] M.A.Kashem et al, “Internet of Things (IOT) based ECG System for Rural Health Care,” Intl. J of Adv Computer Science and Applications , vol. 12, no. 6, pp. 470-477, 2021.

[5] M. A. Serhani, H. T. El Kassabi, H. Ismail, and A. N. Navaz, “ECG monitoring systems: Review, architecture, processes, and key challenges,” Sensors (Switzerland), vol. 20, no. 6, 2020, doi: 10.3390/s20061796.

[6] Z. Yang, Q. Zhou, L. Lei, K. Zheng, and W. Xiang, “An IoT-cloud Based Wearable ECG Monitoring System for Smart Healthcare,” J. Med. Syst., vol. 40, no. 12, 2016, doi: 10.1007/s10916-016-0644-9.

[7] M. Haghi Kashani, M. Madanipour, M. Nikravan, P. Asghari, and E. Mahdipour, “A systematic review of IoT in healthcare: Applications, techniques, and trends,” J. Netw. Comput. Appl., vol. 192, no. January, p. 103164, 2021, doi: 10.1016/j.jnca.2021.103164.

[8] L. R. Yeh et al., “Integrating ECG monitoring and classification via IoT and deep neural networks,” Biosensors, vol. 11, no. 6, pp. 1–12, 2021, doi: 10.3390/bios11060188.

[9] F. Miao, Y. Cheng, Y. He, Q. He, and Y. Li, “A wearable context-aware ECG monitoring system integrated with built-in kinematic sensors of the

smartphone," Sensors (Switzerland), vol. 15, no. 5, pp. 11465–11484, 2015, doi: 10.3390/s150511465.

[10] E. S. Winokur, M. K. Delano, and C. G. Sodini, "A wearable cardiac monitor for long-term data acquisition and analysis," IEEE Trans. Biomed. Eng., vol. 60, no. 1, pp. 189–192, 2013, doi: 10.1109/TBME.2012.2217958.

[11] F. Aktas, C. Ceken, and Y. E. Erdemli, "IoT-Based Healthcare Framework for Biomedical Applications," J. Med. Biol. Eng., vol. 38, no. 6, pp. 966–979, 2018, doi: 10.1007/s40846-017-0349-7.

[12] P. Kaur, H. S. Saini, and B. Kaur, "Wearable sensors for monitoring vital signs of patients," Int. J. Eng. Technol., vol. 7, no. 2, pp. 62–65, 2018, doi: 10.14419/ijet.v7i2.11.11009.

[13] P. Singh and A. Jasuja, "IoT based low-cost distant patient ECG monitoring system," Proceeding - IEEE Int. Conf. Comput. Commun. Autom. ICCCA 2017, vol. 2017-January, pp. 1330–1334, 2017, doi: 10.1109/CCAA.2017.8230003.

[14] U. Satija, B. Ramkumar, and S. M. Manikandan, "Real-Time Signal Quality-Aware ECG Telemetry System for IoT-Based Health Care Monitoring," IEEE Internet Things J., vol. 4, no. 3, pp. 815–823, 2017, doi: 10.1109/JIOT.2017.2670022.

[15] U. Gogate and J. Bakal, "Healthcare monitoring system based on wireless sensor network for cardiac patients," Biomed. Pharmacol. J., vol. 11, no. 3, pp. 1681–1688, 2018, doi: 10.13005/bpj/1537.

[16] N. Clark, E. Sandor, C. Walden, I. S. Ahn, and Y. Lu, "A wearable ECG monitoring system for real-time arrhythmia detection," Midwest Symp. Circuits Syst., vol. 2018-August, pp. 787–790, 2019, doi: 10.1109/MWSCAS.2018.8624097.

[17] P. Kaur, R. Kumar, and M. Kumar, "A healthcare monitoring system using random forest and internet of things (IoT)," Multimed. Tools Appl., vol. 78, no. 14, pp. 19905–19916, 2019, doi: 10.1007/s11042-019-7327-8.

[18] M. Bansal and B. Gandhi, "IoT Big Data in Smart Healthcare (ECG Monitoring)," Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, no. 2, pp. 390–396, 2019, doi: 10.1109/COMITCon.2019.8862197.

[19] N. Huda, S. Khan, R. Abid, S. B. Shuvo, M. M. Labib, and T. Hasan, "A Low-cost, Low-energy Wearable ECG System with Cloud-Based Arrhythmia Detection," 2020 IEEE Reg. 10 Symp. TENSYMP 2020, pp. 1840–1843, 2020, doi: 10.1109/TENSYMP50017.2020.9230619.

[20] F. Ishtiaque, S. R. Sadid, M. S. Kabir, S. O. Ahalam, and M. S. I. Wadud, "IoT-Based Low-cost Remote Patient Monitoring and Management system with Deep Learning-Based Arrhythmia and Pneumonia detection," 2021 IEEE 4th Int. Conf. Comput. Power Commun. Technol. GUCON 2021, pp. 4–9, 2021, doi: 10.1109/GUCON50781.2021.9573620.

[21] M. L. Sahu, M. Atulkar, M. K. Ahirwal, and A. Ahamad, "IoT-enabled cloud-based real-time remote ECG monitoring system," J. Med. Eng. Technol., vol. 45, no. 6, pp. 473–485, 2021, doi: 10.1080/03091902.2021.1921870.

[22] E. A. Ghafil, H. Ghassan, H. Abdulmajeed, and M. H. Majeed, "Remote Cardiac Patients Monitoring System Using Internet of Medical Things (IoMT) Devices," Central Asian Journal of Theoretical and Applied Sciences, vol 3, no. 5, pp. 531–536, May 2022. https://doi.org/10.17605/OSF.IO/SEUYB.

[23] Z. I. Communications, S. Ma, M. Alkhaleefah, Y. Chang, and J. H. Chuah, "Inter-Multilevel Super-Orthogonal Space – Time Coding Scheme," pp. 1–15, 2022.

[24] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG heartbeat classification: A deep transferable representation," Proc. - 2018 IEEE Int. Conf. Healthc. Informatics, ICHI 2018, pp. 443–444, 2018, doi: 10.1109/ICHI.2018.00092.

[25] Q. H. Abbasi, A. A. Nasir, K. Yang, K. A. Qaraqe, and A. Alomainy, "Cooperative In-Vivo Nano-Network Communication at Terahertz Frequencies," IEEE Access, vol. 5, pp. 8642–8647, 2017, doi: 10.1109/ACCESS.2017.2677498.

A Proposed Deep Learning based Framework for Arabic Text Classification

Mostafa Sayed[1]
Faculty of Computers and Artificial Intelligence, Beni-Suef University
Beni-Suef, Egypt

Hatem Abdelkader[2]
Faculty of Computers and Information, Menoufia University
Menoufia, Egypt

Ayman E. Khedr[3]
Information Systems Department
Faculty of Computers and Information Technology, Future University in Egypt (FUE)
Cairo, Egypt

Rashed Salem[4]
Faculty of Computers and Information, Menoufia University
Menoufia, Egypt

***Abstract*—Deep learning has become one of the crucial trends in the modern era due to the huge amount of data that has become available. This paper aims to investigate and improve a generic framework for Arabic Text Classification (ATC) with different deep learning techniques. Besides, it deals directly with a word in its original style as a basic unit of modern Arabic sentence and on a different level of N-grams versus a combination of Intersected Consecutive Word proposed method (ICW). However, it aimed to discuss the results of the different experiments for the enhancements of the proposed method on different deep learning algorithms such as Scaled Conjugate Gradient (SCG) and Gradient descent with momentum and adaptive learning rate backpropagation (GDX) on ATC. The results showed that the proposed framework applied with the SCG algorithm and TF-IDF outperforms the GDX algorithm with an accuracy ratio of 90.65%.**

***Keywords*—Text classification; arabic text classification; scaled conjugate gradient; TF-IDF; GDX; ICW**

I. INTRODUCTION

The classification of texts is becoming more crucial every day due to the tremendous diversity in the use of different human knowledge sources. This usage of cognitive resources resulted in the momentum and abundance of information and data circulating between many devices with large volumes, rapid and remarkable development of artificial intelligence. Therefore, it was necessary to work effectively in containing this momentum in order to classify these texts effectively, not only to facilitate the retrieval of information but also for machine learning uses.

Text classification (TC – also known as text categorization, or topic spotting) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set. This assignment falls at the intersection of Information Retrieval (IR) and Machine learning (ML). It can be defined as the process of classifying or structuring documents into a predefined set of categories according to a group structure that is known in advance [1]. Also, Khorsheed defined it as "The assignment of free-text documents to one or more predefined categories based on their content" [2].

Valuable information can be elicited out of organized/unorganized textual resources like documents and classified into a stable number of predefined categories that are established in advance, they express the fundamental idea of text categorization. Each document can be in multiple, solo, or no class at all [3],[4].

The problems of classifying texts are not only represented in the tremendous diversity in the use of cognitive sources or in the momentum but also the abundance of information and circulating data, which sometimes reach millions of terabytes. Moreover, there is a crucial factor which is the language in which these texts are written [28].

In the field of Natural Language Processing (NLP), different languages were interested in researches development more than the others. Whatever, some of these researches are concerned with the Arabic language as it has a gorgeous impact even it became one of the most commonly used languages all over the world; despite it considers the fifth spoken one. It uses profusely in many of the different Arab countries as it is the main language of the Holy Quran. Moreover, various applications are still limited for the Arabic language owing to its enormous variation in shape, structure, and component, although different studies were carried out for text classification using the English language [5], [26], [27].

There are many challenges concerning the documents written in the Arabic language. From these challenges the common characteristic of the language, e.g., richness of vocabulary, the complexity of grammar, combinations of orthography, the existence of short vowels, …, etc. These problems are particularized in detail in [6], [7]. Besides, the algorithms that were developed for English perform unwell for Arabic [8].

Different researchers interested in the field of NLP are concerned with applying and studying deep learning algorithms in order to explore many results for the reasons of development and improvement [31]. Deep learning is a

coherent and integrated set of algorithms that interpret and link data to each other in order to achieve the greatest degree of accuracy in order to identify and extract new information that was previously unknown [33]. However, the method of learning these algorithms is a representation of the way human brain cells work in transmitting and interacting signals.

Machine learning techniques can be classified for ATC under two categories. The first category is the classical machine learning techniques which contain approved algorithms such as SVM, KNN, NB, and others. Whereas, the second category, is modern machines learning techniques which contain algorithms such as stochastic gradient descent (SGD), convolutional neural network (CNN), and bi-directional long-short term memory (BLSTM).

The contribution of this paper is presenting a novel framework for handling the binary classification problem in Arabic text by employing a new proposed ICW method as a feature representation. Also, the proposed framework handles the effects of deep learning techniques in binary classification problem in Arabic text.

The rest of the paper is organized as follows: The second section mentions the Arabic text classification phases. The third section mentions a literature study of deep learning algorithms and the related works in ATC. The fourth section, introduces some crucial background to facilitate the understanding the following sections. Section five presents the proposed framework based on deep learning techniques. Section six presents the experimental study and discusses the experimental results obtained. Finally, the seventh section concludes the paper's contributions and future work.

II. Arabic Text Classification Phases

Building a generic framework depends on previous phases of text classification. Fig. 1 shows the generic framework extracted based upon [9], [10] was able to take a step forward towards applying deep learning techniques in ATC. Six phases are considered as the main phases for achieving the text classification and dimensionality curse problems.

A. Data Collection Phase

According to the huge usage of the Internet and social media, there are different types of data which are differed in shape and volume. For that, data can be collected through the Internet or detected system in a represented shape such as text, documents, web pages, videos, spreadsheets, or database files. Data can be retrieved through different resources which can be categorized under defined symmetric or asymmetric data. Also, it can be categorized into three categories i.e., structured, unstructured, and semi-structured in the same group.

B. Pre-Processing Phase

Different steps are made for cleaning and preparing the collected data as a result of the last preprocessing phase. Preparing the different data collected from different resources in a homogeneity manner is a prerequisite objective for the classifiers in the classification phase [30]. However, there are general steps used in the Arabic text classification phase such as excluding stop words which include pronouns, conjunctions, and prepositions. Also, exclude digits, punctuation marks, Latin alphabet, removal of isolated letters, and non-Arabic words.

The text prepared before is tokenized and divided by representing it in different manners under the conditions of usage purposes. There are two popular models for the reason of tokenizing text N-gram and bag of words (BOW) models.

C. Representing Phase

After preparing the data, it is represented through an indexed matrix vector space. Indexing is a crucial process in (IR) systems [25], [29]. It reduces the documents into the informative terms contained in them.

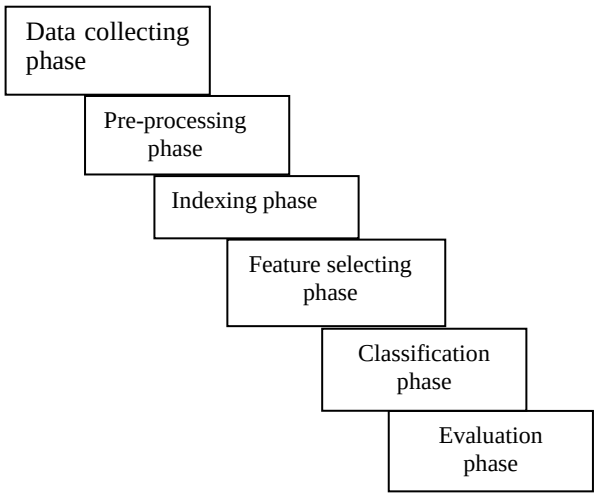


Fig. 1. Conceptual View of Text Classification Steps.

It provides a mapping from the terms to the respective documents containing them [11]. The produced indexed matrix-vector can be represented as equation number (1).

$$M = X * Y \quad (1)$$

Where $X = (D_0, D_1, D_2, \ldots\ldots, D_{n-1}, D_n)$ where D_n represent documents on data set

$Y = (w_0, w_1, w_2, \ldots\ldots, w_{n-1}, w_n)$ where w_n represent the words in the document and $Y \subset D_n$.

When training a machine or a deep learning classification model on a dataset, a matrix of vector space is prepared considerably. The rows of the matrix represent the actual documents that were contained in the training dataset, and the columns represent the features represented in each document.

D. Feature Selecting Phase

In this phase, a matrix of vector space is represented among all extracted features of documents represented in the data set. A massive number of features extracted through the data-set crash with the “Curse of dimensionality” problem. Features reduction algorithms are adaptable techniques for solving that problem that reduces the low priority of features in consideration of the high quality of the text classification process.

E. Classification Phase

It's the important phase through all the previous processes. In this phase, the classifier is trained as a model based on previous train data-set. The objective of this phase is to train the classifier to generalize the method of classifying data on another symmetric data-set.

F. Evaluation Phase

The last phase is evaluating the generated classification model from the previous phases. A symmetric test data-set is prepared for the purpose of measuring the accuracy of the classifier according to the confusion matrix.

III. Related Work

Different studies concerned with machine learning techniques were interested in defining Neural Networks as a biased term for complicated deep learning techniques. CNNs are very similar to ordinary Neural Networks, they are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, executes a dot product, and may follow it with a non-linear function. The entire system still expresses a single differentiable score function: from the inputs to the classes in the outputs. Moreover, they still have a loss function on the last layer. CNN is typically a sequence of layers, and the outputs of every layer are the inputs of the next layer. Those layers are convolutional, pooling layers, and fully connected layers [12], [13].

Sagheer et al, discussed applying deep learning techniques for Arabic sentence classification. They presented a CNN model to achieve the purpose of classifying the text in the used dataset into three labels. The used dataset contained the Arabic sentence withdraw from Essex Arabic Summaries Corpus (EASC). They used different techniques for preparing text for the deep learning algorithm, they first tokenized the words of text and then transformed them to sequences represented by word indices. According to the variation of the sequence length of words, they used a padding technique to trim the sequence length into a unified length. They used the word embedding layer for comparing three models of CNN. For avoiding the overfitting issues through the training phase of the algorithm, they used dropout layers and weighting regularization functions. The results showed that CNN models in top of word embedding layer achieved high-performance accuracy in the NLP task, Arabic sentences classification [14].

In other work, Helmy et al., proposed a deep learning-based approach for keyphrase extraction of Arabic text. As mentioned before a shortage in Arabic text datasets was found for adapting deep learning models for Arabic keyphrase extraction. However, they established a new dataset that was prepared to consist of 6,000 abstracts of scientific Arabic documents. They apply a word embedding representation method for representing the tokens of the documents which are divided into sentences. The embedding layer works as a lookup table that transforms discrete features such as the words of Arabic text into continuous real-valued vector representations, which are then concatenated and provided to the neural network. Also, they proposed applying the BLSTM network for utilization instead of a feed-forward network. They used two hidden layers forward and backward hidden sequence to generate the output. Then, a cunctation layer is connected to a softmax output layer with three neurons for each word. A dropout technique was implemented between Bi-LSTM and the dense layer to prevent overfitting. The evaluation results showed that the proposed approach achieves state-of-the-art performance in the Arabic KPE domain [15].

Also, Samir Boukil et al., proposed a method for Arabic text classification. The proposed method that followed depends mainly on the known steps for preparing the Arabic text as preprocessing steps. Also, they used a stemmer for purpose of extracting and selecting the features. They used TF and TF-IDF techniques as feature weighting techniques for representing the text. For the classification phase, they compared three classifiers CNN, SVM, and Logistic Regression. In the CNN experiment, they employed stochastic gradient descent (SGD) to train the network and use a backpropagation algorithm to calculate the gradients. Besides, they used a learning rate of 0.001 and a dropout ratio with a value of 0.5 to enhance the classifier performance. They argued that the CNN algorithm achieved high results in large and big datasets versus the traditional algorithms [16].

Whatever, deep learning can be a coherent and integrated set of algorithms that interpret and link data to each other in order to achieve the greatest degree of accuracy in order to identify and extract new information that was previously unknown. The method of learning these algorithms is a representation of the way human brain cells work in transmitting and interacting signals. Because of the lack of research in the field of ATC, the previous methodologies used in the mentioned research are considered important points. Also, these researches are characterized as the basic steps in the direction of building generally proposed frameworks to improve deep learning algorithms to classify texts in the Arabic language. Moreover, as mentioned in the literature review several deep learning techniques differ between them in architectures and performance and not all of them have been applied to the ATC problem.

The previous issues have motivated us to propose a large and accessible benchmark dataset of binary-label Arabic texts classified under a legal text. Besides, it has motivated us for exploring in a comparative manner the effect between two deep learning algorithms, i.e., GDX and SCG for ATC.

IV. Background

A. Data Preprocessing based Techniques

In the following, we demonstrate the proposed combination of words for representing the Arabic text based on the structure of the Arabic sentence. Arabic sentence contains in its normal structure two types of sentences, i.e., noun sentence and verbal sentence. The noun sentence consists of two parts or tokens "mobtada" and "khaber" whereas the verbal sentence consists of three parts or tokens subject, verb, and object. For extracting all possible sentences from the text the next two definitions were formulated for that purpose.

Definition 1: Arabic text can be represented by T= (w_0, w_1, w_2, …, w_{n-1}, w_n) Noun Sentence (NS) can represented by NS =

(w_0+w_1,w_1+w_2,……., w_{i-1}+w_i) where w_{i-1}+w_i ⊂ T and w_{i-1}+w_i ≤ w_{n-1}+w_n and i=n .

According to the last definition, several noun sentences NS will be represented by several tokens each one consisting of (wi-1+wi) Ɛ T.

Definition 2: Arabic text can be represented by T= (w_0, w_1,w_2,w_3…,w_{n-2},w_{n-1},w_n) Verbal Sentence (VS) can represented by VS = (w_0+w_1+w_2,w_1+w_2+w_3,……., w_{i-2}+w_{i-1}+w_i) where w_{i-2}+w_{i-1}+w_i ⊂ T and w_{i-2}+w_{i-1}+w_i ≤ w_{n-2}+w_{n-1}+w_n and i=n .

ICW Proposed method

Input: file of text contains a number of lines

Output: separated files each one contains one of a line of words equals to the token value

1: For each (line in lines)

2: If the line is not empty then

3: Read the words in each line

4: For each line do

5: Read the token value

6: Divide the words equals to token value

7: Create a separated file for each line;

8: Write to file the words equals to token value;

9: Exit

B. Feature Selection Phase

The feature selection phase is crucial in our proposed framework as it's the last step in drawing the features vector space for each class. However, two methods are proposed for applying to establish the features vectors are TF and TF-IDF [32]. Term frequency is concerned with how frequently a word or a combination of words occurs in a detected one document. Where TF-IDF is concerned with how frequently a word or a combination of words occurs within the overall document [28].

For calculating TF-IDF assuming that a given a group of documents D, a word w, and an individual document d ϵ D, we calculate *wd* the weight of the word w by applying (1) and (2).

$$wd = TF * IDF \quad (2)$$

$$wd = fwd * log(|D|/fwd) \quad (3)$$

Where fwd is the number of times the word w appears in the document d, |D| is the size of the corpus, and fwD is the number of documents in which w appears in D [17].

C. Classification Phase

In this phase, the algorithm is trained on the data-set for purpose of achieving the classification task. Whatever the well-defined deep learning classification algorithms we choose SCG and GDX.

1) Scaled conjugate gradient back propagation algorithm: This algorithm was developed by (Moller,1990). It was built based upon a network training function that updates weight and bias values according to the scaled conjugate gradient method. It depends on conjugate directions, though it does not perform a line search at each iteration for avoiding the time-consuming linear search of conjugate and optimal direction that occurs with other algorithms.

The scaled conjugate gradient method relies on fast strategy search using information from the second-order approximation [18]. The mathematical equations used for that algorithm can be summarized as follows:

$$E(w+y) \approx E(w) + E'(w)Ty + 1/2yTE(w) \quad (4)$$

The quadratic approximation to E for the point w can be achieved through Eqw(y) in eq (5)

$$Eqw(y) = E(w) + E'(w)Ty + 1/2yTE^{"}(w)y \quad (5)$$

For determining minima to Eqw(y) the critical points must be detected in equation (6). The critical points are the primitive keys for linear systems [19].

$$Eqw(y) = E^{"}(w)y + E'(w) = 0 \quad (6)$$

Assume that conjugate systems with start point Y_1, and $P1…PN$. We can consider a linear combination of the points from Y_1 to Y^* till PN. Where Y^* is a critical point.

$$Y^* - Y1 = \sum_{i=1}^{n} \propto ipi \; where \;\; \propto_1 \in R \quad (7)$$

$$P_j^T\left(-E'(w) - E^{"}(W)y1\right) = \propto jP_j^T E^{"}(w)P_j \quad (8)$$

$$\propto = \left(P_j^T\left(-E'(w) - E^{"}(w)y1\right)\right)/\left(jP_j^T E^{"}(w)P_j\right) \quad (9)$$

Using Eqs (7), (8), and (9), we can iteratively determine the value of the critical point which is Y*. Eqw(Y), is given by the equation,

$$E_{qw}(Y) = E_{qw}(Y^*) + 1/2(Y - Y^*)TE^N(W)(Y - Y^*) \quad (10)$$

2) Gradient descent with momentum and adaptive learning rate back propagation: The algorithm can be defined as it updates weight and bias values according to gradient descent momentum and an adaptive learning rate. Momentum factors can be accomplished by adding a fraction of the previous weight change to the current weight change. This term encourages movement in the same direction on successive steps. The addition of such a term can help smooth out the descent path by preventing extreme changes in the gradient due to local anomalies. Therefore, it is likely to suppress any oscillations that result from changes in the slope of the error surface [20].

Back propagation is used to calculate derivatives of performance *perf* with respect to the weight and bias variables X. Each variable is adjusted according to gradient descent with momentum,

$$dX = mc * dXprev + lr * mc * dperf/dX \quad (11)$$

where, *dXprev* is the previous change to the weight or bias.

D. Evaluation Measures

Precision and recall are widely used for evaluation measures in information retrieval and machine learning [21]. Precision is the fraction of retrieved documents that are relevant to the query. In other words, it concerns how useful the search results are. Recall is the segment of the documents which is exactly related to the inquiry that is absolutely recalled. However, it concerns with how complete results are. F-measure is approximately the average of precision and recall.

TABLE I. DOCUMENTS POSSIBLE SETS BASED ON A QUERY IN IR

Iteration	Relevant	Irrelevant
Documents Retrieved	true positives (tp)	false positives (fp)
Documents not Retrieved	false negatives (fn)	true negatives (tn)

According to Table I, precision, recall, and (macro average) measures can be computed as the following equations:

$$Precision = \frac{tp}{(tp+fp)} \quad (12)$$

$$Recall = \frac{tp}{(tp+fn)} \quad (13)$$

$$F - measure = \frac{2*Precision*Recall}{Recall+Precision} \quad (14)$$

V. PROPOSED FRAMEWORK BASED ON DEEP LEARNING TECHNIQUES

For building a generic framework, it builds based upon the conceptual view of text classification. Fig. 2 shows the generic framework which extracted based upon [9], [10]. Moreover, it is able to take a step forward towards for applying deep learning techniques in ATC. In addition, a new legal dataset for requests of prosecutors was presented for the purpose of testing and training the proposed framework in different experiments.

In the data collection phase, data from two courts of the council state of Egypt are collected and reviewed by three technical reviewers. The reviewing process was performed according to the spatiality of each document of the previously detected two classes.

Different steps have been made for cleaning and preparing the collected data in preprocessing phase such as excluding stop words which include pronouns, conjunctions, and prepositions. Also, exclude digits, punctuation marks, Latin alphabet, removal of isolated letters, and non-Arabic words [24].

Besides, the adding point of the proposed method for representing the text is in the features representation phase. The proposed method was built based upon definition [1, 2] and the ICW algorithm. Two combination forms for representing the text were proposed for evaluating the proposed framework. Firstly, it evaluates the combination between the uni-word and noun sentence. Secondly, it evaluates the combination between noun sentences and verbal sentences.

A comparison between two representation methods in the features representation phase showed. On one hand, the N-gram model with three aspects of a word-level uni-gram, bi-gram, and tri-gram is established. On the other hand, the proposed method for representing the text with its two combinations of uni-word, noun sentence, and verbal sentence is presented.

In classification phase two deep learning algorithms are applied for the purpose of classifying data into its class SCG and GDX. They applied for the purpose of measuring their accuracy with the previous steps of preparing the data of two classes with TF representing features vector in one side. In the other side, they also applied for measuring their accuracy in classification with TF-IDF. Besides, deep learning architectures changed in a manner for evaluating and exploring the effects of changing the number of both hidden learning layers and neurons of each layer in the accuracy of the classifiers. Fig. 2 shows the different steps from collecting the data from documents to the evaluation step of the classifier.

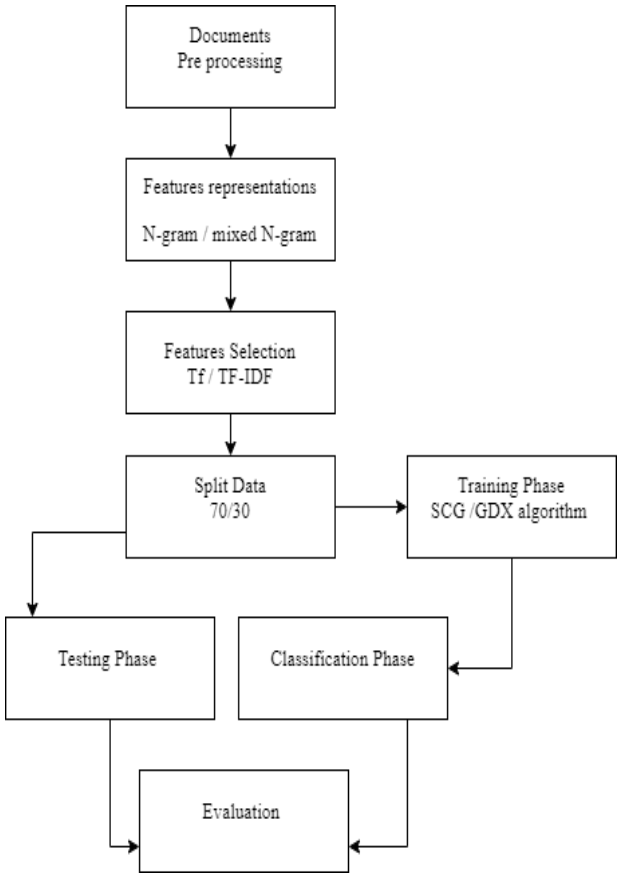


Fig. 2. Proposed Framework for ATC with Deep Learning Techniques.

VI. EXPERIMENTS AND RESULTS

A Matlab software with deep learning toolbox version 2018b was used for implementing different steps of the proposed framework of Fig. 2. Besides, the hardware machine used for implementing the different experiments was a server machine with a specification of Xeon processor 1.96 and memory 32 gigabytes.

A. Dataset

To evaluate any categorization system, text collection which consists of different categories must be available for training and testing purposes. There is no standard Arabic text collection as a benchmark dataset to the best of our knowledge for researchers who work on classifying Arabic text in the legal environment [22], [23]. Wherefore, one of the main purposes of this study is to contribute to the research community by representing a legal dataset written with Arabic text. Its value stems from the real addition to the trends of digital transformation and digitization of the work of the Egyptian judiciary. In addition to establishing the principle of complete justice, transparency, and facilitating the work of the judiciary, within the framework of Egypt's plan for sustainable development 2030.

We collected a subset of the Arabic text for requests of Prosecutors covering two topics from an Arabic dataset built locally at the council state of Egypt. The reasons for choosing these Prosecutors' requests are firstly, each Prosecutors request has a unique label. This makes it easier for us to access and fetch the files by their label. Secondly, each request has been reviewed by two different phases of technical reviewing.

Two categories were collected of this dataset containing 500 text documents for each class the total size is 1000 text documents. The total features extracted from these data exceed thousands of words beginning from a least 2000 words and in some cases exceed 16000 features. For more details of the dataset, description sees Table II.

The following statement is an example of ATC problem which labeled as class one:

اولا: قبول الدعوى شكلا. ثانيا: وفي الموضوع الحكم بالزام المدعى عليهم بصفتهم بان يؤدوا للطالبه اجرا مضاعفا عن عملها ايام الراحات الاسبوعية والعطلات والاجازات والبالغ قدرها 326 يوم ثلاثمائة وستة وعشرون يوما مع الزام الجهة الادارية بالمصروفات والاتعاب على ان ينفذ الحكم بمسودته دون اعلان.

Moreover, the following statement is labeled as class two.

اولا: قبول الدعوى شكلا. ثانيا: وفى الموضوع الحكم بالزام المدعى عليه بان يؤدى للطالب اجر مضاعف عن جميع ايام الراحات الاسبوعية والعطلات الرسمية واجازات الاعياد منذ التحاقه بالعمل حتى الان مع الزام المدعى عليه بالمصروفات والاتعاب مع حفظ كافة حقوق الطالب الاخرى.

Different preprocessing steps have been made for handling the last data such as removing the duplicated words and digits etc.…another example for tokenizing the text into the different methods such as N-gram or ICW method is described below.

Example one: tokenizing the text into a bi-gram method.

باحقية المدعى # المدعى المعاملة # المعاملة المالية # المالية طبقا # طبقا لقرار # لقرار رئيس # رئيس مجلس # مجلس الوزراء # الوزراء رقم # رقم لسنة # لسنة فترة # فترة ابتعاثة # ابتعاثة لدولة

Example two: tokenizing the text into a tri-gram method.

باحقية المدعى المعاملة # المدعى المعاملة المالية # المعاملة المالية طبقا # المالية طبقا لقرار # طبقا لقرار رئيس # لقرار رئيس مجلس # رئيس مجلس الوزراء # مجلس الوزراء رقم # الوزراء رقم لسنة # رقم لسنة فترة # لسنة فترة ابتعاثة # فترة ابتعاثة لدولة

TABLE II. DATASET DESCRIPTION ACCORDING TO N-GRAM AND PHRASE STRUCTURE

	Features of Class one	**Features of Class two**	**Before remove duplicate**	**After remove duplicate**
uni-gram	1331	1489	2820	2182
bi-gram	3826	3734	7560	6831
tri-gram	5220	4910	10430	9582
one word and noun phrase	5115	5198	10313	8952
noun phrase and verbal phrase	8997	8613	17610	16339
one word, noun phrase and verbal phrase	10286	10077	20363	18460

B. Experimental Configuration

The experimental configuration was built based on illustrating the effect of changing the learning layers architecture for the algorithm on the classifier's accuracy. Also, it was built based on illustrating the effect of using the proposed method for representing Arabic text based on noun and verbal sentences with various combinations versus N-gram. Different types of experiments with detailed sub-experiments were configured for achieving the last two objectives. First, experiments were set up for comparing the accuracy of the classifier between representing the uni-gram with TF and representing the uni-gram with TF-IDF with the increasing number of layers respecting two a constant of the number of neurons in each layer. Second, experiments were set up for comparing the accuracy of the previous deep learning algorithm represented in SCG algorithm with (TF) and (TF-IDF) features selection method with a proposed ICW method of a uni-bi gram and bi-tri gram.

Other experiments were set up for comparing the accuracy of the SCG algorithm versus the GDX algorithm with a two and three hidden layer with 100 neurons for each layer. However, according to the main objective of the comparisons (TF-IDF) and TF features selection method is still used for unifying the configuration of the experiments.

C. Evaluating the Accuracy of SCG Algorithm through Changing the Architecture Layers:

The experiment was executed based upon the framework in Fig. 2 with the detailed steps which have been discussed before. This experiment was established for the purpose of exploring the effects of changing the architecture layers in the accuracy level of the classifier. The first factor is increasing the number of learning hidden layers whereas the other factor is fixing the number of neurons. We measure the accuracy level of the classifier with equation number based on the confusion matrix and equation (14). Comparisons have been made between representing the uni-gram with TF and the uni-gram with TF-IDF with the increasing number of layers.

$$Layer\ size = nl * N \quad (15)$$

Where nl >= 2 and N is a constant equal 100.

The factor nl is the number of layers and N in the number of neurons.

The experiments are executed till the stopping condition is achieved. The stopping condition is the accuracy level is lower than the first experiment with the number of layers being 2.

TABLE III. ACCURACY OF SCG ALGORITHM WITH CHANGING THE ARCHITECTURE LAYERS

Number of layers	Uni-gram TF	Uni-gram TF-IDF
LayerSize = [2*100]	88.02	86.29
Layer Size = [3*100]	88.74	87.18
Layer Size = [4*100]	88.77	89.67
Layer Size = [5*100]	88.53	88.69
Layer Size = [6*100]	88.08	86.90
Layer Size = [7*100]	88.89	87.21
Layer Size = [10*100]	86.98	84.85
Layer Size = [20*100]	89.61	87.41
Layer Size = [30*100]	89.19	90.07
Layer Size = [50*100]	87.12	86.86
Layer Size = [80*100]	85.1	86.11

The results in Table III showed that increasing the number of hidden layers with a fixed number of neurons affects the accuracy of the classifier with TF and TF-IDF representation methods. Also, Fig. 3 showed a variance of accuracy ratios changed with increasing the number of hidden layers. However, it has been detected that the SCG algorithm achieved the best accuracy ratio with 90.07% with the number of hidden layers equaling 30 layers.

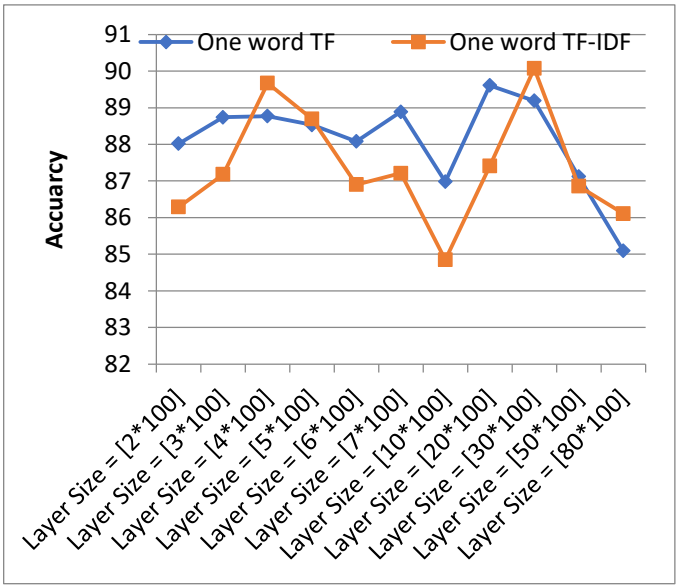


Fig. 3. The Effect of Accuracy According to the Change of Layers Number with Fixed Neurons.

D. Accuracy Ratios of SCG Algorithm According to TF and TF-IDF Features Selection with Fixed Layers and Neurons:

The experiment was executed based upon the framework in Fig. 2 with the detailed steps that have been discussed before. This experiment was established for the purpose of exploring the effects of two factors in the accuracy level of the classifier. The first factor is a fixed number of learning hidden layers whereas the other factor is a fixed number of neurons for each layer. Table II showed the number of words represented as features used in this experiment. We measure the accuracy level of the equation number based on the confusion matrix. Comparisons have been made between two different methods on one hand a combination between (uni-gram and bi-gram) and (bi-gram and tri-gram) on the other hand the typical N-gram model. The last two combinations are compared with TF and TF-IDF representing methods with a fixed number of layers. The fixed number of layers is represented with an assumed number of neurons which equals 100 neurons for each one of the layers.

Table IV shows the results for discussing the factors that affects the accuracy of the classifier with both representation methods TF and TF-IDF. The main purpose of the experiment stills the highest accuracy for text classification that achieved with two detected layers and 100 of neurons. It has been showed that representing the text with TF-IDF representation method outperform TF representation method in bi-gram, tri-gram, and (uni-bi) gram with ratios 90.34%,88.36%, and 90.65% respectively.

TABLE IV. ACCURACY OF SCG ALGORITHM ACCORDING TO TF AND TF-IDF FEATURES SELECTION WITH FIXED LAYERS AND NEURONS

	Tf	Tf-idf
Uni-gram	88.02	86.29
Bi-gram	89.78	90.34
Tri-gram	87.79	88.36
(Uni,Bi)-gram	89.55	90.65
(Bi,Tri)-gram	88.76	88.55

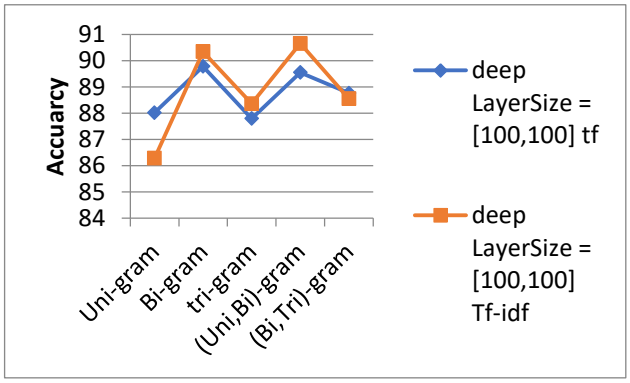


Fig. 4. The Change of Accuracy for Proposed Method versus N-Gram Model.

Moreover, the best result achieved absolutely in this experiment was in the proposed combination of (uni-bi) gram with ratio 90.65%. Fig. 4 shows the highest of accuracy ratios between the N-gram and the proposed combination of words with TF and TF-IDF representation methods.

E. Comparing of the SCG and the GDX According ICW Proposed Method Versus N-Gram Method

The experiment was executed based upon the framework in Fig. 2 with the detailed steps which have been discussed before. This experiment was established to explore the effects of both fixed number of hidden learning layers and fixed number of neurons for each layer in the accuracy level of the classifier. Besides, it compares different combination methods such as uni-bi gram, bi-tri gram versus the N-gram with different representation methods. We measure the accuracy level of the equation number based on the confusion matrix.

Tables V and VI shows the results for discussing the factors that affect the accuracy of the classifier with both representation methods TF and TF-IDF. The main purpose of the experiment stills the highest accuracy for text classification that achieved with a detected number of layers and 100 neurons between the proposed combination method (ICW) and the N-gram. It has been shown that the proposed combination uni-bi gram with the SCG algorithm outperforms the other technique with three layers and 100 neurons even with the TF representation method or with the TF-IDF representation method with ratios of 89.45% and 89.60%, respectively.

TABLE V. COMPARING OF THE SCG VERSUS THE GDX ACCORDING N-GRAM METHOD

Feature selections	Layer size	Algorithm	Uni-gram	Bi-gram	tri-gram
			ACC	ACC	ACC
Tf	Ls[2*100]	scg	88.02	89.78	87.79
		gdx	85.82	82.23	80.12
	Ls[3*100]	scg	88.74	88.35	87.86
		gdx	87.35	81.98	79.52
Tf-idf	Ls[2*100]	scg	86.29	90.34	88.36
		gdx	86.07	81.95	79.68
	Ls[3*100]	scg	87.18	87.86	87.55
		gdx	84.22	83.94	79.90

TABLE VI. COMPARING OF THE SCG VERSUS THE GDX ACCORDING ICW PROPOSED METHOD

Feature selections	Layer size	Algorithm	(Uni,Bi)-gram	(Bi,Tri)-gram
			ACC	ACC
Tf	Ls[2*100]	scg	89.55	88.76
		gdx	84.93	80.52
	Ls[3*100]	scg	89.45	86.62
		gdx	83.93	83.20
Tf-idf	Ls[2*100]	scg	90.65	88.55
		gdx	83.62	80.27
	Ls[3*100]	scg	89.60	87.66
		gdx	83.99	80.31

Moreover, the best result achieved absolutely in this experiment was in the proposed combination of (uni-bi) gram with a ratio of 90.65% with the TF-IDF representation method, two layers, and 100 neurons. In addition, the last comparison mentioned that the SCG algorithm outperforms the GDX algorithm in all experiments.

VII. CONCLUSION

This paper aims to investigate and to improve a generic framework for Arabic text classification. It deals directly with a word in its original style as a basic unit of modern Arabic sentence and on different levels of N-grams versus ICW proposed method. However, it aimed at discussing the results of the different experiments for studying the effect of changing the architecture concerning learning layers of different deep learning algorithms on ATC as a case study. In addition, a new legal dataset for requests of Prosecutors was presented for the purpose of testing and training the proposed framework in different experiments.

The main results that are drawn from this work showed that with increasing the number of hidden layers with a fixed number of neurons affects the accuracy of the SCG classifier with TF and TF-IDF representation methods respecting to uni-gram. However, it has been detected that the SCG algorithm achieved the best accuracy ratio with 90.07% with the TF-IDF representation method; also, in comparing the accuracy of the SCG algorithm between N-gram and the proposed method (ICW) with a fixed number of layers and neurons. It has been shown that representing the text with the TF-IDF representation method and the proposed method (ICW) (uni-bi) gram outperforms TF representation method with ratio 90.65%. Moreover, it has been shown that the proposed method (ICW) (uni-bi) gram with the SCG algorithm with TF-IDF representation method outperforms the GDX algorithm with a ratio of 90.65%.

For future work, the proposed model needs to be tested with different large datasets as a benchmark for generalizing and extracting a lot of results. Besides, the proposed model needs to integrate an optimization technique as feature reduction for enhancing the "curse of dimensionality" problem.

REFERENCES

[1] Khreisat, L., A machine learning approach for Arabic text classification using N-gram frequency statistics, Journal of Informatics, Volume 3, 2009.

[2] Khorsheed, Mohammad S., and Abdulmohsen O. Al-Thubaity, Comparative evaluation of text classification techniques using a large diverse Arabic dataset, Language Resources and Evaluation, 2013.

[3] Joachims.T, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proceedings of the European Conference on Machine Learning (ECML), Springer, 1998.

[4] Elhassan.R, Ahmed.M, Arabic Text Classification on Full Word, International Journal of Computer Science and Software Engineering (IJCSSE), Volume 4, Issue 5, 2015.

[5] Abd El-Monsef .M. E, Amin.M, Atlam.E, El-Barbary.O, Arabic Document Classification: A Comparative Study,Journal of Computing, Volume 3 ,Issue 4, April 2011.

[6] Abbès, R., Dichy, J, AraConc, An Arabic Concordance Software Based on the DIINAR.1 Language Resource. In: The 6th International Conference on Informatics and Systems, 2008.

[7] Darwish. K., Building a shallow Arabic morphological analyzer in one day. In Proceedings of the ACL 2002 Workshop on Computational Approaches to Semitic Languages, Stroudsburg, PA, USA, pages 1–12, 2002.

[8] Abuaiadah. D, Arabic Document Classification Using Multiword Features , International Journal of Computer and Communication Engineering, Vol. 2, No. 6, 2013.

[9] F. Al-Zaghoul and S. Al-Dhaheri, Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks, Proceedings of the 15th International Conference on Computer Modelling and Simulation (UKSim), Cambridge University, United Kingdom, , pp. 485-490, April 2013 .

[10] Sayed. M, Salem. R and Khedr. AE, A survey of Arabic text classification approaches, International Journal of Computer Applications in Technology 59(3):236 – 251, March 2019 .

[11] H. Kaur and V. Gupta, Indexing process insight and evaluation, International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1-5, doi: 10.1109/INVENTIVE.2016.7830087.

[12] A. Krizhevsky , I. Sutskever and G. E. Hinton, ImageNet Classification with Deep Convolution Neural Networks, Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS'12), Lake Tahoe, Nevada, (2012) December, vol 1, pp. 1097-1105.

[13] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, Gradient-based learning applied to document recognition", Proceedings of the IEEE, vol. 86, no. 11, (1998) November, pp. 2278-2324.

[14] Dania Sagheer, Fadel Sukkar, "Arabic Sentences Classification via Deep Learning ", International Journal of Computer Applications (0975 – 8887) Volume 182 – No.5, July 2018.

[15] Muhammad Helmy, R. M. Vigneshram , et al ," Applying Deep Learning for Arabic Keyphrase Extraction "The 4th International Conference on Arabic Computational Linguistics (ACLing 2018), 2018, Dubai, United Arab Emirates.

[16] Samir Boukil, Mohamed Biniz, Fatiha El Adnani3, Loubna Cherrat and Abd Elmajid El Moutaouakkil. Arabic Text Classification Using Deep Learning Technics. International Journal of Grid and Distributed Computing Vol. 11, No. 9 (2018), pp.103-114.

[17] G. Salton and C. Buckley, "Term-weighing approache in automatic text retrieval", Information Processing & Management Journal., vol. 24, no. 5, (1988), pp 513-523.

[18] M.F. Moller.: A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning, Neural Networks, Vol. 6, pp. 525–533, 1993.

[19] Maryam Habibie and Andrei Propescu-Belis.: Keyword Extraction and Clustering for Document Recommendation in Conversations, IEEE/acm Transactions on Audio, Speech, and Language Processing, vol. 23, no. 4, pp. 746, April 2015.

[20] M.Z. Rehman and N.M. Nawi : ICSECS 2011, Springer ,Part I, CCIS 179, pp. 380–390, 2011.

[21] Thabtah, F., Eljinini, M., Zamzeer, M., & Hadi, W. Naı¨ve Bayesian based on Chi Square to categorize Arabic data. In Proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, pp. 930–935, 2009.

[22] Alharbi, F.R., Khan, M.B. Identifying comparative opinions in Arabic text in social media using machine learning techniques. SN Appl. Sci. 1, 213, 2019.

[23] Al-Taani A.T., Al-Sayadi S.H. Classification of Arabic Text Using Singular Value Decomposition and Fuzzy C-Means Algorithms. In: Johri P., Verma J., Paul S. (eds) Applications of Machine Learning. Algorithms for Intelligent Systems. Springer, Singapore,2020.

[24] El-Fishawy, N., Hamouda, A., Attiya, G. M., & Atef, M. Arabic summarization in Twitter Social Network. Ain Shams Engineering Journal, 5(2), 411–420. https://doi.org/10.1016/j.asej.2013.11.002, 2014.

[25] M Attia, MA Abdel-Fattah, Ayman E. Khedr," A proposed multi criteria indexing and ranking model for documents and web pages on large scale data ". Journal of King Saud University-Computer and Information Sciences,2021.

[26] Boudad, N., Faizi, R., Oulad Haj Thami, R., & Chiheb, R. Sentiment Analysis in Arabic: A review of the literature. Ain Shams Engineering Journal, 9(4), 2479–2490. https://doi.org/10.1016/j.asej.2017.04.007, 2018.

[27] Al-Anzi, F. S., & AbuZeina, D. Synopsis on Arabic speech recognition. Ain Shams Engineering Journal, 13(2), 101534. https://doi.org/10.1016/j.asej.2021.06.020, 2021.

[28] Amira M. Idrees, Essam M. Shaaban," Building a Knowledge Base Shell Based on Exploring Text Semantic Relations from Arabic Text". International Journal of Intelligent Engineering and Systems, Vol.13, No.1, 2020.

[29] Mohamed, M. A., Abdel-Fattah, M. A., & Khedr, A. E. (2021). Challenges and Recommendations in Big Data Indexing Strategies. International Journal of e-Collaboration (IJeC), 17(2), 22-39. http://doi.org/10.4018/IJeC.2021040102

[30] Khedr, A. E., Idrees, A. M., & Alsheref, F. K. (2019). A Proposed Framework to Explore Semantic Relations for Learning Process Management. International Journal of e-Collaboration (IJeC), 15(4), 46-70. http://doi.org/10.4018/IJeC.2019100104.

[31] Mostafa, A. M., Idrees, A. M., Khedr, A. E., & Helmy, Y. M. (2020). A proposed architectural framework for generating personalized users' query response. Journal of Southwest Jiaotong University, 55(5). https://doi.org/10.35741/issn.0258-2724.55.5.3

[32] Amr Mansour Mohsen, Hesham Ahmed Hassan, Amira M. Idrees.(2016). Documents Emotions Classification Model Based on TF-IDF Weighting Measure. International Journal of Computer and Information Engineering Vol:10, No:1.

[33] Afify, E. A., Sharaf, A., & Khedr, A. E. (2020). Facebook profile credibility detection using machine and deep learning techniques based on user's sentiment response on status message. International Journal of Advanced Computer Science and Applications, 11(12). https://doi.org/10.14569/ijacsa.2020.0111273

Simultaneous Importance-Performance Analysis based on SWOT in the Service Domain of Electronic-based Government Systems

Tenia Wahyuningrum[1], Aina Azalea[4]
Department of Informatics
Institut Teknologi Telkom
Purwokerto, Banyumas, Indonesia

Gita Fadila Fitriana[2], Arief Rais Bahtiar[3]
Department of Software Engineering
Institut Teknologi Telkom
Purwokerto, Banyumas, Indonesia

Darwan[5]
Department of Mathematics Education
IAIN Syekh Nurjati
Cirebon, Indonesia

***Abstract*—Decision makers for decades have used SWOT analysis for strategic planning. However, the problems that arise in the SWOT analysis are subjective, so decision-making becomes inefficient. Therefore, SWOT analysis is often combined with other methods to make decision-making strategies more focused and measurable according to priority interests. The SWOT analysis basis in this study is Simultaneous Importance-Performance (SIPA) analysis by observing each indicator's weights. In addition, this study proposes a new method by focusing on competitor factors in strategies mapping to improve services for Electronic-Based Government Systems (SPBE). The object of this study was two local governments in Indonesia, namely the Meranti Islands Regency and the Limapuluh Kota Regency. The results showed that a SIPA-based SWOT analysis has succeeded in showing the Strengths, Weaknesses, Opportunities, and Challenges of the district government. Furthermore, based on the results of hypothesis testing, SIPA-based SWOT identification has reflected a valid organizational situation.**

Keywords*—Importance performance analysis; strength weakness opportunity threat analysis; service quality; electronic based government systems

I. Introduction

For approximately 60 years, Strengths, Weak, Opportunities, and Threat analysis (abbreviated as SWOT) have been a key and fundamental tool in strategic planning [1]. This analysis evaluates the organization's position to see its position in its internal and external environment. Strategic planning generally uses SWOT analysis, but the method is subjective and only focuses on solving weaknesses separately [2]–[4]. This problem is because the basis for SWOT analysis traditional approach is a qualitative analysis where SWOT factors tend to have a subjective view on the assessment of managers and planners, so they are considered inefficient and lead to wrong business decisions [2]. To prove its validity and accuracy, researchers often combine SWOT analysis with other techniques in various problems solving such as educational, industrial, agricultural, environmental, and economic [5]. The combination of SWOT analysis with other methods such as Analytical Hierarchy Process (AHP), Fuzzy AHP, Analytic Network Process (ANP), and Importance-Performance Analysis (IPA) shows that SWOT analysis is a flexible model [2], [6]. Based on several combinations of these methods, the combination of the SWOT-IPA method is considered accurate and valid to describe the organizational situation. Initially, managers used the IPA method as a marketing tool. However, its application extended to various fields, such as tourism, teaching, food service, health care, money-saving, human resources, and data innovation [7]–[13]. We find that each indicator has equal weight in some of these areas. Nevertheless, in some instances, each measure may have a different weight and need to be compared with other research objects to see the difference.

The combined IPA and SWOT methods [14] have not involved the problem for each indicator's initial weight. However, the indicator's weight determines a value's importance and performance. In addition, the IPA method only considers internal organizational aspects and dismisses the company's external factors. One of the improvements to the IPA method, namely Simultaneous Importance-Performance Analysis (SIPA), is a modification of the IPA method that map the relationship between the importance and performance of product/service quality attributes [9], [13], [15]–[17]. The modification made by SIPA is to pay attention to competitors' aspects in an organization's analysis [18]. In order to reduce these two deficiencies, this study applied SIPA to identify SWOT based on an SPBE survey conducted by the central government and local governments (self-assessment). In order to evaluate the SPBE services in Indonesia, the government has formulated each factor's weight [19].

The authors hope that by using SWOT-based SIPA analysis, organizations (in this case, local governments) can formulate strategic planning efficiently because the SWOT factors that must be maintained and improved can be identified based on the community's point of view. This study compares two local governments, namely the Meranti Islands Regency and the Limapuluh Kota Regency. These two regencies' location is on the island of Sumatra, Indonesia, with almost the same area and population. However, the Meranti Islands Regency was only established in 2008 and is a division of the Bengkalis Regency. Therefore, to accelerate the implementation of e-government, it is necessary to map various indicators of strengths, weaknesses, opportunities, and challenges. This paper consists of six sections: Introduction,

literature review, proposed method, result, discussion, and conclusions.

II. Literature Review

A. SPBE Service Indicators

SPBE services, according to the Regulation of the Ministry of State Apparatus Utilization and Bureaucratic Reform of the Republic of Indonesia Number 59 of 2020 concerning Monitoring and Evaluation of Electronic-Based Government Systems, consist of measuring the organization's service capability [19]. Therefore, the indicators used to measure the maturity level of SPBE services are electronic-based government administration services and electronic-based public services. Furthermore, there are five levels of measurement for service capability maturity: information, interaction, transactions, collaboration, and optimum. Therefore, the indicators listed in the SPBE service domain can be described as the aspects of electronic-based government administration services and aspects of electronic-based public services as follows:

The aspect of Electronic-Based Government Administration Services, with the weight of each indicator (ω_i), is 0.0604.

a) Indicator 1 Maturity Level of Planning Services.

b) Indicator 2 Maturity Level of Budgeting Service.

c) Indicator 3 Maturity Level of Financial Services.

d) Indicator 4 Maturity Level of Procurement Services.

e) Indicator 5 Maturity Level of Personnel Services.

f) Indicator 6 Maturity Level of Dynamic Archival Services.

g) Indicator 7 Maturity Level of State/Regional Property Management Services.

h) Indicator 8 Maturity Level of Government Internal Supervision Services.

i) Indicator 9 Maturity Level of Organizational Performance Accountability Services.

j) Indicators of 10 Maturity Level of Employee Performance Service.

The aspect of Electronic-Based Public Services, with the weight of each indicator (ω_i), is 0.0659.

a) Indicator 11 Maturity Level of Public Service Complaints Service.

b) Indicator 12 Maturity Levels of Open Data Services

c) Indicator 13 Maturity Level of Documentation Network and Legal Information

d) Indicator 14 Maturity Level of Public Service Sector 1

e) Indicator 15 Maturity Level of Public Service Sector 2

f) Indicator 16 Maturity Level of Public Service Sector 3.

B. Simultaneous Importance-Performance Analysis (SIPA)

SIPA is a modification of IPA which added competitor factors in evaluating the organization [20]. The description for some of the stages in conducting the SIPA analysis is as follows:

Step 1. Define the weight ω_i of each SPBE service indicator by expert or government regulation, which i is the number of attributes.

Step 2. Collect data through questionnaires performance (χ_{ij}), and importance (γ_{ij}) indicator of item i in the local government j as χ_{ij} and γ_{ij} multiplied with ω_i. Then, calculate the performance and importance as formula (1).

$$\chi_{ij} = \omega_i \cdot \chi_{ij}; \gamma_{ij} = \omega_i \cdot \gamma_{ij} \quad (1)$$

Step 3. The coordinates of SIPA are then divided by equation (2). Then, a judgment on the quadrant SPBE service indicator should be put on each local government (see Fig. 1).

$$\bar{\chi} = \sum_{i=1}^{k}\sum_{j=1}^{l}\chi_{ij} \Big/ K; \bar{\gamma} = \sum_{i=1}^{k}\sum_{j=1}^{l}\gamma_{ij} \Big/ K \quad (2)$$

Fig. 1 shows 4 quadrants, quadrant 1 (top priority), quadrant 2 (keep achievement), quadrant 3 (low priority), and quadrant 4 (excessive).

Step 4. Summarize and categorize the results as below:

If the indicator is in quadrants 1 and 4, it is labeled strength. If the indicator is in quadrants 2 and 3, it is labeled weakness.

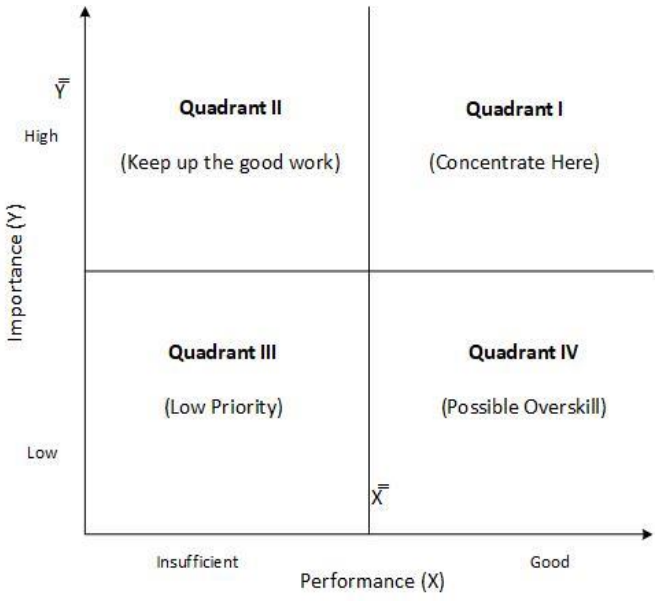


Fig. 1. IPA Matrix [21].

C. Strength, Weakness, Opportunity, and Threat (SWOT) Analysis

The analysis of strengths (S), weaknesses (W), opportunities (O), and threats (T) (SWOT) summarizes the central elements taken by studying the external and internal environment of each organization. Strengths include the organization's internal capabilities, resources, and positive situational factors in achieving its goals. On the other hand,

Weaknesses are internal limitations and negative situational factors that can hinder the organization [3]–[5]. The SWOT table generates the Importance-Performance analysis results, with the researchers' provisions in Table I.

TABLE I. SWOT IDENTIFICATION TABLE

Strength-Weakness		***SWOT Aspect***	***Implication***
Organization	Competitor		
S	S	S	Head-to-head competition
	W	O	Competitive advantage
W	S	T	Competitive disadvantage
	W	W	Neglected opportunities

III. PROPOSED METHOD

The method proposed in this study consists of several stages: filling out questionnaires by the central government (performance) and local governments (importance). Then the second step is to simultaneously conduct an importance and performance analysis on a local government and competitors. Finally, the results of the SIPA analysis become the basis for a SWOT analysis in a local government. Fig. 2 describes the proposed method.

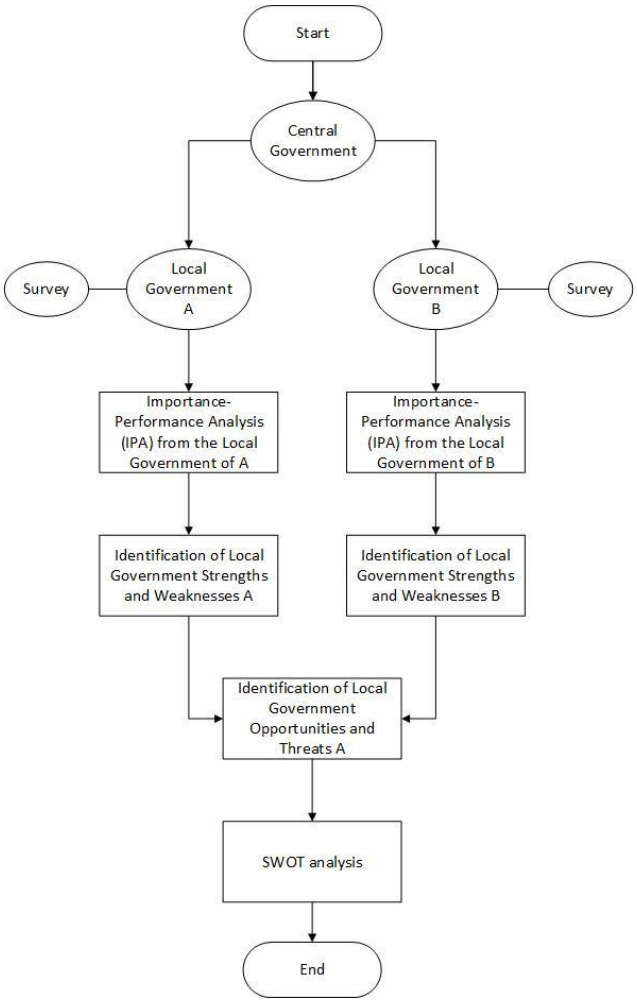


Fig. 2. Proposed Method.

IV. RESULT

A. Case Studies in the Meranti Islands Regency and Limapuluh Kota

This study reviewed two local governments of the Meranti Islands Regency and Limapuluh Kota Regency, Indonesia. These two local governments have almost the same area and population. However, the Meranti Islands local government is still quite young because of the expansion of the Bengkalis Regency. The Meranti Islands Regency has the vision to create good, clean, and responsible governance to provide excellent service, including implementing SPBE. Therefore, the Meranti Islands Regency needs a special strategy to map the appropriate needs and activities to achieve its goals.

B. SIPA Analysis

Based on the results of the SIPA analysis, the importance value was obtained from an independent assessment by Meranti Regency (district A) and Limapuluh Kota Regency (B). Meanwhile, the researchers used a questionnaire to obtain the performance value from a central government assessment through the Ministry of Administrative and Bureaucratic Reform of the Republic of Indonesia. The questionnaire to fill out is on the web https://monev.spbe.go.id/. Each indicator value is multiplied by its weight and entered in the corresponding quadrant. The quadrant determination is as follows:

1) If the performance value (χ_{ij}) is less than the overall average value $\sum \chi_{ij}$, it falls into the insufficient category. On the contrary, if the performance value exceeds the overall average value, it falls into the good category.

2) If the value is importance (γ_{ij})less than the average value of the overall importance $\sum \gamma_{ij}$, then it falls into the low category. On the contrary, the importance value is more than the overall average value, so it falls into the high category.

Table II shows that in District A, the indicators in quadrant 1 (Q1) are 9, 13, and 14. While in quadrant 2 (Q2) is indicator number 1-4, 12, 15-16. In quadrant 3 (Q3), there are indicator numbers 6-8 and 10-11. District B's indicators in quadrant 1 (Q1) are 2, 4, 11, and 15. In Q2, its indicator number is 16; in Q3, the indicator numbers are 6-9 and 13. Finally, in quadrant 4 (Q4), the indicator is 1, 3, 5, 12, 14.

Fig. 3 and Fig. 4 show more detail on the IPA matrix. The pictures show that the value in quadrant 1 is an indicator value with high importance (high) and good performance (good), so the indicators in this quadrant can be maintained. Whereas quadrant 2 shows high importance indicator values (high) but poor performance values (insufficient), it is an indicator that must be aware. Organizational concentration needs to focus on increasing the value of these indicators. Quadrant 3 contains the indicators of importance value that are less important (low) and have poor performance (insufficient), so they are indicators with low priority. Quadrant 4 contains low importance indicators (low) but good performance (good), so the indicators in this quadrant are excessive, allowing them to be the final priority.

TABLE II. IMPORTANCE-PERFORMANCE OF THE TWO DISTRICT

Indicator	District A			District B		
	I	*P*	*IPA Quadrant*	*I*	*P*	*IPA Quadrant*
Indicator 1	0.181	0.060	Q2	0.181	0.242	Q4
Indicator 2	0.181	0.060	Q2	0.242	0.242	Q1
Indicator 3	0.181	0.060	Q2	0.181	0.242	Q4
Indicator 4	0.181	0.060	Q2	0.242	0.242	Q1
Indicator 5	0.181	0.060	Q2	0.181	0.242	Q4
Indicator 6	0.060	0.060	Q3	0.181	0.181	Q3
Indicator 7	0.060	0.060	Q3	0.181	0.181	Q3
Indicator 8	0.060	0.060	Q3	0.181	0.181	Q3
Indicator 9	0.181	0.121	Q1	0.181	0.181	Q3
Indicator 10	0.060	0.060	Q3	0.181	0.181	Q3
Indicator 11	0.066	0.066	Q3	0.264	0.264	Q1
Indicator 12	0.198	0.066	Q2	0.198	0.264	Q4
Indicator 13	0.198	0.132	Q1	0.198	0.132	Q3
Indicator 14	0.198	0.132	Q1	0.198	0.330	Q4
Indicator 15	0.198	0.066	Q2	0.264	0.264	Q1
Indicator 16	0.198	0.066	Q2	0.264	0.198	Q2
Average	**0.149**	**0.074**	-	**0.207**	**0.222**	-

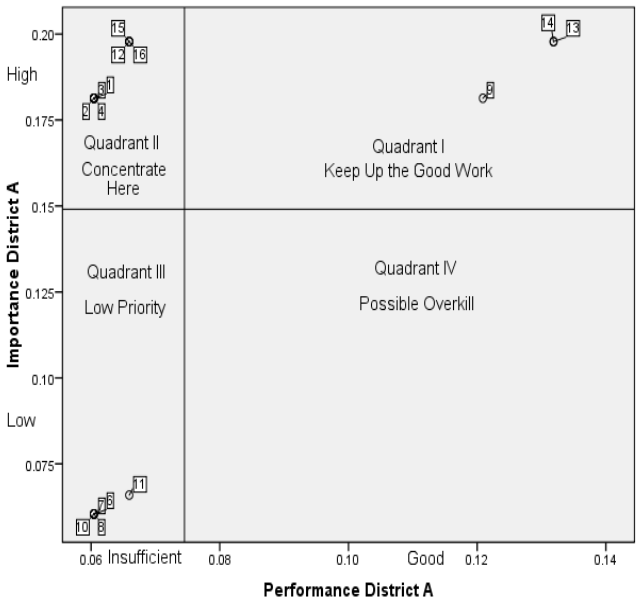


Fig. 3. IPA Quadrant for District A.

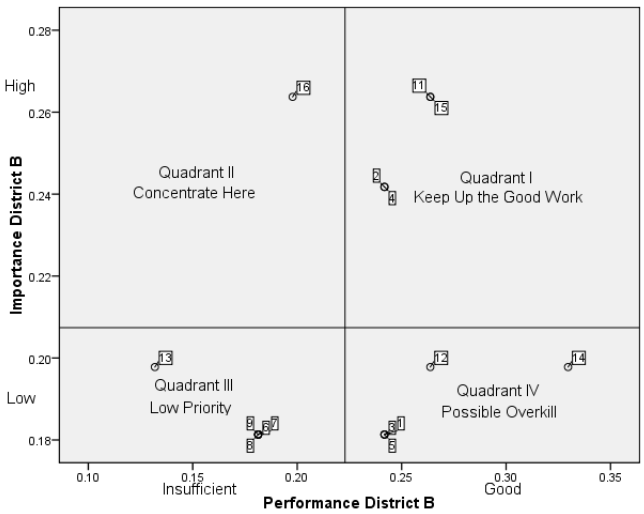


Fig. 4. IPA Quadrant for District B.

Based on the results in Fig. 3 and Fig. 4, the next step is to create the Strength-Weakness table. If the indicator is in quadrants 1 and 4, the label is strengths, and if it is in quadrants 2 and 3, the label is weakness. According to the SWOT identification in Table I, a SWOT analysis was formed for District A. The researchers calculate the aggregate weight value from the importance value multiplied by the final performance. The sign (-) indicates Threat or Weakness, while the sign (+) on the Aggregate Weight indicates Strength or Opportunity. Table III shows the SWOT analysis of District A in detail.

TABLE III. STRENGTH-WEAKNESS OF DISTRICT A

Indicator	*Strenght-Weakness*		*SWOT District A*	*Aggregate weight**
	District A	District B		
Indicator 1	Weakness	Strength	Threat	-0.0109
Indicator 2	Weakness	Strength	Threat	-0.0109
Indicator 3	Weakness	Strength	Threat	-0.0109
Indicator 4	Weakness	Strength	Threat	-0.0109
Indicator 5	Weakness	Strength	Threat	-0.0109
Indicator 6	Weakness	Weakness	Weakness	-0.0036
Indicator 7	Weakness	Weakness	Weakness	-0.0036
Indicator 8	Weakness	Weakness	Weakness	-0.0036
Indicator 9	Strength	Weakness	Opportunity	0.02191
Indicator 10	Weakness	Weakness	Weakness	-0.0036
Indicator 11	Weakness	Strength	Threat	-0.0043
Indicator 12	Weakness	Strength	Threat	-0.0130
Indicator 13	Strength	Weakness	Opportunity	0.02608
Indicator 14	Strength	Strength	Strength	0.02608
Indicator 15	Weakness	Strength	Threat	-0.0130
Indicator 16	Weakness	Weakness	Weakness	-0.0130

[a.] Compute by multiplying the importance by positive/negative performance (for strength, opportunity/weakness, threat)

Table III shows the indicators in the SWOT category in District A,

Strength:

Indicator 14 Maturity Level of Public Service Sector 1.

Opportunity:

1) Indicator 9, Maturity Level of Organizational Performance Accountability Services.
2) Indicator 13, Maturity Level of Documentation Network and Legal Information.

Threat:

1) Indicator 1, Maturity Level of Planning Service.
2) Indicator 2, Maturity Level of Budgeting Service.
3) Indicator 3, Maturity Level of Financial Services.
4) Indicator 4, Maturity Level of Procurement Services.
5) Indicator 5, Maturity Level of Staffing Services.
6) Indicator 11, Maturity Level of Public Service Complaints Service.
7) Indicator 12, Maturity Level of Open Data Services.
8) Maturity Level of Public Service Sector 2.

Weakness:

1) Indicator 6, Maturity Level of Dynamic Archival Service.
2) Indicator 7, Maturity Level of State/Regional Property Management Services.
3) Indicator 8, Maturity Level of Government Internal Oversight Services.
4) Indicator 16, Maturity Level of Public Service Sector 3.

Thus, 8 threat and 4 weakness indicators should be an important concern of stakeholders in determining the priorities of the SPBE implementation strategy in District A.

V. Discussion

We tested the results using a questionnaire to ascertain whether the proposed method has met stakeholders' satisfaction. We used another questionnaire of 6 District A staff respondents to clarify these findings. The survey is done for evaluation purposes since there is no direct method or tool to validate the effectiveness of an IPA-based SWOT analysis [2]. The evaluation questionnaire consisted of closed questions asking the respondents' approval level in District A for the results of the IPA-based SWOT analysis shown in Table IV. Each question uses a score of a four-point Likert scale (1 = Strongly disagree to 4 = Strongly agree) without a midpoint that acts as a neutral choice. The hypothesis formulated is as follows:

H_0: The mean response is equal to 2.5.

H_a: The mean response is not equal to 2.5.

The tested hypothesis is at a significance level of 5%, with one sample t-test analysis shown in Table IV.

TABLE IV. Strength-Weakness of District A

Variable	*Mean*	*SD*	*Mean difference*	*t*	*df*	*Sig. (2-tailed)*
W_{avg}	3.08	0.10	0.58	12.124	3	0.001
O_{avg}	3.58	0.12	1.08	13	1	0.049
T_{avg}	3.25	0.31	0.75	6.87	7	0.000

Table IV shows that the average respondent's weakness, opportunity, and threat assessments are at 3.08, 3.58, and 3.25. Whether to accept or reject the hypothesis shown from the calculated *t* values on weakness, opportunity, and threat, respectively, namely 12.124, 13, and 6.87, with degrees of freedom 3, 1, and 7. This data shows that the table t values for weakness, opportunity, and threat are 3.182, 12.71, and 2.365.

The decision-making is done by comparing the calculated t-values and t of the table. When observed, the calculated t values for weakness, opportunity and threat are greater than t of the table; this means the results reject H_0. Similarly, when viewed from the significance values of the three variables below 0.05, it also rejects H_0. This result means that the respondent (District A staff) agreed with the IPA-based SWOT results' Strengths, Weaknesses, Opportunities, and Threats. Based on the analysis, this model is recommended for decision-making by considering the weight of the criteria and competitor factors.

VI. Conclusion

This research results from implementing a SIPA-based SWOT analysis that measures the level of importance as a representation of expectations from the organization to performance assessed by other parties. This method also considers the weights on each of the indicators and also the competitors of an organization. Taking into account internal and external factors, shows that district A has 8 indicators of threat and 4 indicators of weakness out of 16 indicators of electronic-based government system services. In this analysis, respondents confirmed and approved the results regarding the Strengths, Weaknesses, Opportunities, and Threats of District A on the SPBE service indicator. Thus, this model can be used for decision-making by considering the weights of indicators and competitor factors in various cases. The future work in this research is to combine the SWOT method with other methods, such as the Simple Additive Weighting (SAW), Technique for Order by Similarity to Ideal Solution (TOPSIS), Profile Matching (PM), and other appropriate procedures. Further development can also be focused on the number of additional research objects.

Acknowledgment

Thank you to LPPM Institut Teknologi Telkom Purwokerto for funding this research with IT contract number Tel2411/LPPM-000/Ka. LPPM/IV/2022 on behalf of Tenia Wahyuningrum. Thank you to Mr. Amat Safii, head of the communication and information section of the Meranti Islands Regency, and Mr. Fery Chofa, head of the communication and information service of Limapuluh Kota Regency, Indonesia, who have been willing to become the object of research.

REFERENCES

[1] M. A. Benzaghta, A. Elwalda, M. M. Mousa, I. Erkan, and M. Rahman, "SWOT analysis applications : An integrative literature review," J. Glob. Bus. Insights, vol. 6, no. 1, pp. 55–73, 2021, doi: 10.5038/2640-6489.6.1.1148.

[2] B. Phadermrod, R. M. Crowder, and G. B. Wills, "Importance-Performance Analysis based SWOT analysis," Int. J. Inf. Manage., vol. 44, pp. 194–203, 2019, doi: 10.1016/j.ijinfomgt.2016.03.009.

[3] R. Madurai Elavarasan, S. Afridhis, R. R. Vijayaraghavan, U. Subramaniam, and M. Nurunnabi, "SWOT analysis: A framework for comprehensive evaluation of drivers and barriers for renewable energy development in significant countries," Energy Reports, vol. 6, pp. 1838–1864, 2020, doi: 10.1016/j.egyr.2020.07.007.

[4] C. Vlados, "On a correlative and evolutionary SWOT analysis," J. Strateg. Manag., vol. 12, no. 3, pp. 347–363, 2019, doi: 10.1108/JSMA-02-2019-0026.

[5] A. Adem, A. Çolak, and M. Da, "An integrated model using SWOT analysis and Hesitant fuzzy linguistic term set for evaluation occupational safety risks in life cycle of wind turbine," vol. 106, no. May 2017, pp. 184–190, 2018, doi: 10.1016/j.ssci.2018.02.033.

[6] R. G. Dyson, "Strategic development and SWOT analysis at the University of Warwick," vol. 152, pp. 631–640, 2004, doi: 10.1016/S0377-2217(03)00062-6.

[7] M. Lettner, F. Hesser, B. Hedeler, and P. Schwarzbauer, "Barriers and incentives for the use of lignin-based resins: Results of a comparative importance performance analysis," J. Clean. Prod., vol. 256, no. 5, p. 120520, 2020, doi: 10.1016/j.jclepro.2020.120520.

[8] M. Cladera, "An application of importance-performance analysis to students ' evaluation of teaching," Educ. Assessment, Eval. Account., vol. 33, pp. 701–715, 2021, doi: 10.1007/s11092-020-09338-4.

[9] J. J. Kim, Y. Lee, and H. Han, "Exploring competitive hotel selection attributes among guests : An importance-performance analysis," J. Travel Tour. Mark., vol. 36, no. 9, pp. 998–1011, 2019, doi: 10.1080/10548408.2019.1683484.

[10] O. A. Ogunmokun, K. Kolawole, T. Avci, T. Temitope, and J. E. Ikhide, "Propensity to trust and knowledge sharing behavior : An evaluation of importance-performance analysis among Nigerian restaurant employees," Tour. Manag. Perspect., vol. 33, no. May 2019, p. 100590, 2020, doi: 10.1016/j.tmp.2019.100590.

[11] S. F. Kak and F. M. Mustafa, "Smart Home Management System Based on Face Recognition Index in Real-Time," in 2019 International Conference on Advanced Science and Engineering, ICOASE 2019, 2019, pp. 40–45, doi: 10.1109/ICOASE.2019.8723673.

[12] H. Mustafa, B. Omar, and S. N. S. Mukhiar, "Measuring destination competitiveness: an importance-performance analysis (IPA) of six top island destinations in South East Asia," Asia Pacific J. Tour. Res., vol. 25, no. 3, pp. 223–243, 2020, doi: 10.1080/10941665.2019.1687534.

[13] J. Bi, Y. Liu, Z. Fan, and J. Zhang, "Wisdom of crowds : Conducting importance-performance analysis (IPA) through online reviews," Tour. Manag., vol. 70, no. September 2018, pp. 460–478, 2019, doi: 10.1016/j.tourman.2018.09.010.

[14] B. Phadermrod, R. M. Crowder, and G. B. Wills, "Importance-Performance Analysis based SWOT analysis," Int. J. Inf. Manage., vol. 44, pp. 194–203, 2019, doi: 10.1016/j.ijinfomgt.2016.03.009.

[15] N. Ummi, N. Wahyuni, and I. Apriadi, "Analysis of Service Quality on Customer Satisfaction Through Importance Performance Analysis and KANO Model," J. Ind. Serv., vol. 6, no. 2, pp. 1–9, 2021, doi: 10.36055/62013.

[16] J. Hua and W. Y. Chen, "Prioritizing urban rivers ' ecosystem services : An importance-performance analysis," Cities, vol. 94, no. May, pp. 11–23, 2019, doi: 10.1016/j.cities.2019.05.014.

[17] L. Xie, Y. Chen, B. Xia, and C. Hua, "Importance-Performance Analysis of Prefabricated Building Sustainability : A Case Study of Guangzhou," Adv. Civ. Eng., vol. 2020, pp. 1–16, 2020, doi: 10.1155/2020/8839118.

[18] Y. Lee and Y. Hsieh, "Integration of revised simultaneous importance performance analysis and decision making trial and evaluation laboratory: A study of the mobile telecommunication industry in Taiwan," Glob. J. if Bus. Manag., vol. 5, no. 6, pp. 2312–2321, 2017, doi: 10.5897/AJBM10.979.

[19] Menteri PAN RB, "Pemantauan dan Evaluasi Sistem Pemerintahan Berbasis Elektronik." Jakarta, Indonesia, pp. 1–59, 2020.

[20] Y. C. Lee, Y. F. Hsieh, and C. W. Huang, "Using Gap Analysis and Implicit Importance to Modify SIPA," in International Conference on Industrial Engineering and Engineering Management, 2010, pp. 175–179, doi: 10.1109/ICIEEM.2010.5646639.

[21] J. Kwon and T. Chung, "Importance-Performance Analysis (IPA) of Service Quality for Virtual Reality Golf Center," Int. J. Mark. Stud., vol. 10, no. 3, pp. 30–40, 2018, doi: 10.5539/ijms.v10n3p30.

Federated Learning and its Applications for Security and Communication

Hafiz M. Asif[1]
Department of Electrical & Computer Engineering
Sultan Qaboos University
Muscat, Oman

Mohamed Abdul Karim[2]
Department of Information Technology
University of Technology and Applied Sciences
Suhar Campus, Oman

Firdous Kausar[3]
Department of Electrical & Computer Engineering
Sultan Qaboos University
Muscat, Oman

***Abstract*—The not so long ago, Artificial Intelligence (AI) has revolutionized our life by giving rise to the idea of self-learning in different environments. Amongst its different variants, Federated Learning (FL) is a novel approach that relies on decentralized communication data and its associated training. While reducing the amount of data acquired from users, federated learning derives the benefits of popular machine learning techniques, it brings learning to the edge or directly on-device. FL, frequently referred to as a new dawn in AI, is still in its early stages and is yet to garner widespread acceptance, owing to its (unknown) security and privacy implications. In this paper, we give an illustrative explanation of FL techniques, communication, and applications with privacy as well as security issues. According to our findings, there are fewer privacy-specific dangers linked with FL than security threats. We conclude the paper with the challenges of FL with special emphases on security.**

Keywords*—Federated learning; communication; security; deep learning; Artificial Intelligence

I. Introduction

In the modern era, ubiquitous mobile gadgets are coupled with computation and sensor capabilities that collect large volumes of data. Such massive quantities of data are used to train various learning algorithms. These learning techniques, when combined with Data Mining and AI in other words with Deep Learning (DL) breakthroughs, enable a wide range of beneficial applications, including image analysis, speaker identification, healthcare, vehicular networks, among others. Machine Learning (ML) techniques need to be checked in and generally have to be consolidated on internet-based cloud services. However, due to the enormous volumes of data and privacy-critical nature, login into such cloud services to train supervised learning is cumbersome. As a result, major challenges such as excessive latency and transmission inefficiencies arise. The notion of Federated Training or Learning (FL) has now been proposed in face of emerging privacy rules in various nations. Mobile phone users in Federated Learning (FL) can train a feature map by pooling their native models without disclosing their confidential material. In ML, a model is usually developed by training locally on the user's own server (PC) whereas, in FL, the model is built by training on different machines located at distributed locations and there is no dedicated connection between the servers. They have their dataset or database sample at their ends. In simple terms, a form of machine learning that is decentralized is termed federated learning [1]. While there has been some research on this subject, there is not enough progress in terms of comprehending FL's security and privacy implications. This paper aims to provide a full review of FL in terms of a formal definition, then we compare ML models with salient features and tabulate. We also discuss the pros and cons graphically along with the challenges. Finally, we provide some recommendations, making this unique among previous studies. The following is a summary of this paper's contributions to the field's recent literature:

- Providing a categorization and review of the FL methods and strategies.
- Identifying and examining pros and cons in FL environments.
- Delineating potential applications of FL environments.
- Highlighting challenges faced by FL systems with special emphasis on the security.
- Providing recommendations to enhance the security and privacy of the FL implementation.

II. Federated Model and Critical Analysis

ML techniques traditionally require that all the training data be centralized on a single server in a datacenter or the cloud. With enormous increase in the number of mobile devices and the training data available on various machines, the challenge of assimilating the relevant data arises. Federated learning uses the model training approach that enables a device to train from the collaboration of shared models. Proxy data on the server initially trains the shared model. Subsequently, the model is downloaded on each device and then improved by data locally stored on the device, which is also termed as federated data. ML algorithms assume that all learning data is available and maintained in a centralized dataset. In order to facilitate such training, centralized learning networks are created. These networks have serious privacy concerns, high communication costs, and scalability challenges. Federated Learning (FL) has been introduced to enable remote supervised learning without a centralized training classifier, considering the aforementioned difficulties. It can be observed from Fig. 1(b) that the federated learning network is composed of multiple Edge Devices (EDs)

and servers. According to a survey, in federated learning networks, a machine-learning model is trained with two iterative steps such as local model training and global model aggregation at EDs and server respectively [2]. In the first stage, EDs update the local model with the downloaded model from the server, algorithms of Stochastic Gradient Descent are executed to learn the local model with their dataset and upload the updated model to the server. In the next phase, model updates received by the server are aggregated with weighted average to previous global model and thus the new model is obtained. These two steps involve a training round. In a federated learning model or network, the parameters of the ML model are exchanged instead of data and this prevents and reduces privacy issues with the reduction of communication overhead. Federated learning is deployed with flexibility in multiple environments including the mobile environment that is a complicated one [3]. Comparison of machine learning models with salient features is given in Table I.

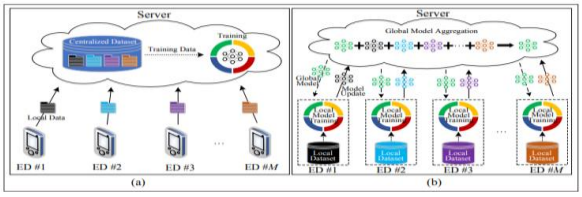


Fig. 1. System Model between Centralized ML Network vs Decentralized FL Network [4].

Federated Learning Network (FLN) has also been adopted to orchestrate various mobile devices across the world for training language models with BERT [7]. All such mobile devices are owned by different users and then connected to multiple types of links such as WiFi, mobile network, etc. Hence, in terms of ownership, capabilities, and computing, Edge Devices (EDs) in federated learning model are heterogeneous [1].

TABLE I. COMPARISON OF MACHINE LEARNING MODELS WITH SALIENT FEATURES [2]

Scheme	Salient features	Used in percentage according to survey
Distributed learning	Provision of holistic estimation of parameters	21%
Parallel learning	Distribution of data in laid fashion	27%
Federated learning	Model training using natural database, massive distribution of data over local learners	45%
Ensemble learning	Production of an optimal model	7%

Federated Learning (FL) is reliable for joint ED's efforts for the training of the ML model. Even with an abnormality of few EDs, Machine Learning model can be tampered. Besides all this, the FL model or network has multiple attack surfaces concerned with the security of federated learning such as malicious EDs, and insecure connections. These attack surfaces are vulnerable to many security issues in FL networks such as data positioning as well as model positioning [4]. Survey analysis of the given graph shows the results of two cases in comparison with using active federated learning framework (Fig. 2).

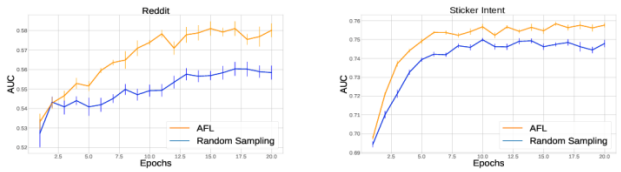


Fig. 2. Comparison of AUC on Reddit and Sticker Intent using Active FL Framework [8].

III. ADVANTAGES OF FEDERATED LEARNING

Diversity of data: Large-scale ML models may be unable to merge datasets from diverse sources. The reasons for impediments are partially due to the information security, reluctance and connection unavailability among the edge devices. On the other hand, Federated learning makes it easier to access diverse data, even when sources of data could only interact at a particular period (Fig. 3).

- Real-time learning continuity: There is no requirement for aggregate data in continuous learning because algorithms are constantly upgraded using client information.
- The efficiency of hardware: Since decentralized learning methods do not require a single, complex cloud database to interpret data, this strategy requires less complex hardware infrastructure [6].

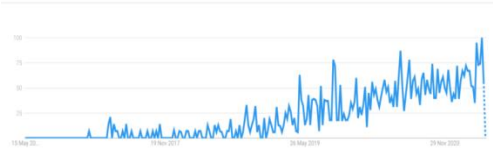


Fig. 3. Interest Overtime Related to use of Federated Learning [5].

IV. APPLICATIONS OF FEDERATED LEARNING

In terms of distributed machine learning, FL is a viable approach with the inherent characteristics of privacy. Without communicating data, many nodes can work together to develop a collaborative learning model. Data access rights, privacy, security, and access to a variety of data types can all be managed in this way. It is believed that FL has applications in a wide range of areas such as Industrial IoT, healthcare, smart transportation, self-driving cars; traffic forecasting; smart buildings, recommender system, Fintech, the insurance sector, and telecommunications [13-14].

FL is a revolutionary technique for machine learning. It has the potential to have a profound impact on the healthcare sector. It can also help healthcare workers in many ways. Sharing health care data raises a number of privacy issues. In addition, strict laws like HIPPA make it more challenging to exchange critical data, which has made it more difficult to conduct studies that could lead to new medical advancements. All parties, including hospitals, AI businesses, and regulatory

authorities, have a responsibility to protect extremely sensitive information. Researchers are currently investigating how FL may be utilized to protect patients' privacy while beneficially utilizing their data. FedHealth [14] is the first federated transfer learning framework for wearable healthcare, capable of providing precise and individualized healthcare without risking patient privacy. A community-based federated learning algorithm (CBFL) [15] proposes a system that clusters distributed data into clinically significant communities based on shared diagnoses and geographical locations and then develops a model for each community. Li et al. [16] develop a brain tumor segmentation FL system using differential privacy to protect patient data. Patients with uncommon tumors will benefit from Owkin's FL-based platform, which will be used in tests to determine drug toxicity, predict disease progression, and assess survival rates [17].

Zhang et al. [18] propose an Industrial Internet identification using blockchain and federated learning technologies, which provides privacy protection. Liu et al. [19] develop an on-device FL-based deep anomaly detection system for IIoT time series data sensing, which detects edge devices' failure in IIoT industrial product production. Edge device failures adversely affect IIoT industrial product production. Khanal et al. [20] examine the value of proactive content caching in self-driving cars to reduce content retrieval costs and improve QoE with edge cloud infrastructure. It extracts local content popularity patterns in self-driving automobiles utilizing LSTM-based prediction mechanisms in a federated scenario to predict regional content popularity.

Machine learning is constantly growing and reshaping the technological landscape. FL applications, like any other machine learning technique, face challenges. In spite of its flaws, it has the potential to transform numerous industries. There will be tremendous progress in FL and its diverse applications soon. When applied effectively, it can aid in the evolution of numerous sectors and benefit users.

Another area where FL finds its rigorous application is data communication. For instance, the feasibility of FL for its using in 6G communication systems has been investigated in [21]. The FL key challenges for 6G include security, cost-effective systems, and privacy concerns. FL can also be used for data augmentation in wireless communication. For instance, edge users can cooperate by sharing certain parameters, which in turn significantly reduces the communication overhead [22]. FL also finds its application in Wireless Power Transfer (WPT) where Wireless-Power enabled can be enabled. A complete wireless-power enabled FL has been investigated in [23].

V. Challenges of Federated Learning with Special Emphases on Security

There are multiple disadvantages related to security issues of the federated learning model. These include data positioning as well as model positioning. The main aim of the positioning attack is to degrade the accuracy of the machine-learning model. This happens by tampering the aggregation of global models with updates of the poisoned model of federated learning [9]. The attack surfaces for such insecurities are Malicious Edge Devices (MEDs) and insecure connections. MEDs are set by the attackers in smart devices through malware. As new smart devices are more sophisticated and have inescapable flaws. Hence, it is convenient for attackers to join the Federated Learning Networks (FLNs) through malicious EDs. Moreover, security of all connections through which the EDs of federated learning model are connected to the network needs be monitored. Wireless connection has vulnerability through various channels. Through such insecure connections, the uploaded model updates of Federated Learning might be manipulated or hijacked. In data poisoning attacks on the security of federated learning, these intentional attacks intend to achieve low accuracy of machine learning models on certain classes. Attackers to the federated learning securities flip labels of training data in those concerned classes [5].

In model poisoning attack on the security of federated learning, the attack is concerned with the ML model updates that are generated from Gaussian distribution (see Table II). In this, the attacker manipulates updates of the benign model into poisoned updates. To achieve this purpose, attackers use updates of pre-model designs to craft the updates of the poisoned model and replace the ML model with the pre-designed poisoned models. The vulnerability points for all these attackers are insecure connections and malicious Edge Devices (EDs). Furthermore, other disadvantages are performance limitations, indirect leakage of information, and a degree of centralization [10].

TABLE II. Attacks at Securities of Federated Learning [8]

Security attacks	**Description**	**Methodology for attacks**	**Target users**
Data poisoning	Training data is modified by attackers and EDs training is made incorrect as well as poisoned updates of the model are generated.	Labels of training data are flipped intentionally, and labels of training data are also flipped in certain classes.	Unintentionally
Model poisoning	Poisoned updates of the model are created by attackers Benign model updates are manipulated based on pre-designed rules.	• Model updates are generated using pre-designed poisoned model to impact the security of federated learning. • Flipping signs of model updates.	Intentionally

Lots of investment is required for federated learning models with frequent communication and large storage capacity with high bandwidth. Data is not collected on a single entity, which increases attack surfaces [1]. The below Fig. 4 depicts the secure aggregation of private federated learning. In this scenario, aggregator or server builds a global model jointly without revealing the security of training data. Hence, it is powerful in terms of keeping privacy while computing millions of data in parallel [3] (Fig. 5).

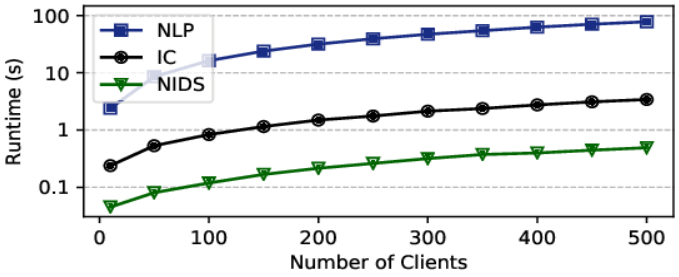


Fig. 4. Secure Aggregation of Private Federated Learning [1].

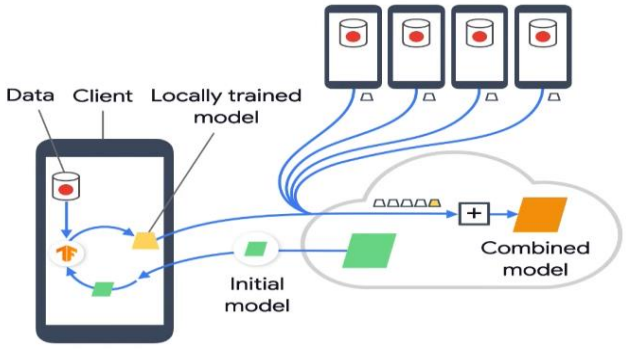


Fig. 5. A Typical Federated Learning Model.

A. Active Federated Learning

The gadgets obtain a training program (which is normally small size in terms of few bytes).

- The gadgets are programmed to learn from local data.
- The sensing notes the computer anonymized updates mostly on variables.
- The data from devices are aggregated by the administrator. The server combines the information it receives from each variety of technologies to conduct an approach with regards to the present system by each grouping.
- The newly added model is delivered to the gadgets with an assessment (again, the idea of decentralization is at work here) as well as a fresh round between training after several rounds of learning [8] (Fig. 6).

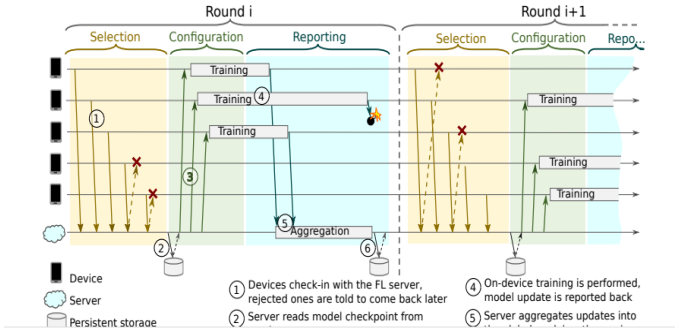


Fig. 6. Federated Learning Protocol [11].

VI. Recommendations and Suggestions

Privacy preservation, safe multiparty processing, and cryptography are examples of confidentiality technologies that can be utilized to improve the data protection possibilities of federated learning. A variety of measures is suggested in this section. First, sharing less information about the generic model updation at the server can maximize the privacy of the Federated-learning model [12]. Moreover, the use of deep neural network also makes complex the use of available gradients. There is also possibility for developers to choose or create an algorithm that has less chance of data breaching and attack on the security of federated learning system. Using more privacy regulations will also inevitably make data acquisition easier and less vulnerable to exploitation [1].

VII. Conclusion

The paper discussed the federated learning techniques and applications with respect to privacy as well as security issues. Federated learning has been successfully implemented in a variety of settings, such as the challenging mobile environment. Despite the advantages of federated learning, there are many privacy and security issues related to the model. When contrasted to exchanging personal data across data centres, federated learning offers certain privacy benefits. The capability to immensely develop machine-learning algorithms depending on user input, while minimizing bandwidth impacts for uploading confidential information over the network is also one of the advantages. Data poisoning and model poisoning are two major security attacks on federating learning networks. Among communication networks, wireless connections are vulnerable. Federated learning system updates can also be altered or hijacked over such unsafe connections. In information poisoning threats on supervised learning of federated model security, these deliberate attacks aim to achieve poor sensitivity of machine learning techniques on specific classes. These attacks are vulnerable through two attack surfaces of federated learning mode such as internet connections and Edge Devices (EDs) of the federated learning model.

Acknowledgment

The authors would like to thank Sultan Qaboos University and University of Technology and Applied Sciences for their support.

REFERENCES

[1] M. Asad, A. Moustafa, T. Ito and M. Aslam, "Evaluating the Communication Efficiency in Federated Learning Algorithms," IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 552-557, 2011.

[2] Uddin, M.P., Xiang, Y., Lu, X., Yearwood J., and Gao L., "Mutual Information Driven Federated Learning," IEEE Transactions on Parallel and Distributed Systems, vol.32, no.7, 1526-1538, 2021.

[3] Asad, M., "Federated Learning Versus Classical Machine Learning: A Convergence Comparison," Journal of scientific research, vol.2, no.2, 23-31, 2019.

[4] McMahan, B., & Ramage, D., "Federated learning: Collaborative machine learning without centralized training data," Google Research Blog, 3, 2017, https://ai.googleblog.com/2017/04/federated-learning-collaborative.html accessed on 11 March 2022.

[5] Dilmegani, C., "What is Federated Learning(FL)?," 2020, Techniques & Benefits in 2021. [Online] https://research.aimultiple.com/federated-learning/ accessed on 11 March 2022.

[6] Tan, J., Liang, Y. –C., Luong, N.C., and Niyato, D., "Toward Smart Security Enhancement of Federated Learning Networks. IEEE Network," vol.35, no.1, pp.340-347, 2021.

[7] lmkil, A., Callh, S., Barbieri, M., S¨utfeld, L.R., Zec, E.L., Mogren, O., "Scaling federated learning for fine-tuning of large language models," Métais, E., Meziane, F., Horacek, H., Kapetanios, E. (eds) Natural Language Processing and Information Systems, 2021.

[8] Jack Goetz, Kshitiz Malik, Duc Bui, Seungwhan Moon, Honglei Liu, Anuj Kumar, "Active Federated Learning," arXiv preprint arXiv:1909.12641, 2019.

[9] Li, T., Sahu, A.K., Talwalkar, A. and Smith, V., "Federated learning: Challenges, methods, and future directions," IEEE Signal Processing Magazine, vol.37, no.3, pp.50-60, 2020.

[10] Fung, C., Yoon, C.J. and Beschastnikh, I., "Mitigating sybils in federated learning poisoning," arXiv preprint arXiv:1808.04866, 2018.

[11] Aledhari, M., Razzak, R., Parizi, R.M. and Saeed, F., "Federated learning: A survey on enabling technologies, protocols, and applications," IEEE Access, 8, 140699-140725, 2020.

[12] Xu, R., Baracaldo, N., Zhou, Y., Anwar, A. and Ludwig, H., "Hybridalpha: An efficient approach for privacy-preserving federated learning," In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, 13-23, 2019.

[13] IEEE Guide for Architectural Framework and Application of Federated Machine Learning. IEEE Std 3652.1-2020,1-69, 2021.

[14] Shaheen, M.; Farooq, M.S.; Umer, T.; Kim, B.-S., "Applications of Federated Learning; Taxonomy, Challenges, and Research Trends," Electronics. Vol.11, 670. https://doi.org/10.3390/electronics11040670, 2022.

[15] Chen Y., Qin X., Wang J., Yu C. and Gao W., "FedHealth: a federated transfer learning framework for wearable healthcare," IEEE Intelligent Systems, vol.35, no.4, pp.83-93, 2020.

[16] W. Li, et al., "Privacy-preserving federated brain tumour segmentation," Machine Learning in Medical Imaging, doi:10.1007/978-3-030-32692-0_16, 2019.

[17] Online, "Federated Learning—OWKIN," Available online: https://owkin.com/federated-learning/ (accessed on June, 2022).

[18] Zhang X., Hou H., Fang Z., and Wang Z., "Industrial Internet Federated Learning Driven by IoT Equipment ID and Blockchain," Wireless Communications and Mobile Computing, Article ID 7705843, 9 pages, 2021.

[19] Liu Y., Garg S., Nie J., Zhang Y., Xiong Z., Kang J., Hossain M., "Deep Anomaly Detection for Time-Series Data in Industrial IoT: A Communication-Efficient On-Device Federated Learning Approach," IEEE Internet of Things Journal, vol.8, no.8, pp.6348-6358, 2021.

[20] Khanal, S., Thar, K., Hossain, M. D., and Huh, E.N., "Proactive Content Caching at Self-Driving Car Using Federated Learning with Edge Cloud," Twelfth International Conference on Ubiquitous and Future Networks (ICUFN), pp.129-134,2021.

[21] Liu, Y., Yuan, X., Xiong, Z., Kang, J., Wang, X., and Niyato, D., "Federated learning for 6G communications: Challenges, methods, and future directions," China Communications, vol.17, no.9, 105-118, 2020.

[22] Yan, M., Chen, B., Feng, G., and Qin, S., "Federated Cooperation and Augmentation for Power Allocation in Decentralized Wireless Networks," IEEE Access, vol.8, pp.48088-48100, 2020.

[23] B. Clerckx, B., Huang, K., Varshney, L. R., Ulukus, S., and Alouini M.S., "Wireless Power Transfer for Future Networks: Signal Processing, Machine Learning, Computing, and Sensing," IEEE Journal of Selected Topics in Signal Processing, vol.15, no.5, 1060-1094. 2021.

Machine Learning in OCR Technology: Performance Analysis of Different OCR Methods for Slide-to-Text Conversion in Lecture Videos

Geeta S Hukkeri, R H Goudar, Prashant Janagond, Pooja S Patil
Department of CSE, VTU
Belagavi, India

***Abstract*—A significant percentage of a lecture video's content shown is text. Video text can therefore be a crucial source for automated video indexing. Researchers have recognised printed and handwritten text extracted from pictures using a variety of machine learning techniques and tools before digitising it. A machine learning technology called optical character recognition (OCR) enables us to recognise and retrieve text information from documents, converting it into searchable and editable data. This study primarily focuses on text extraction from lecture slides using Google Cloud Vision (GCV), Tesseract, Abbyy Finereader, and Transym OCR and compares the results to develop a lecture video indexing scheme for the non-linear steering in lecture videos to watch only the interesting points of topics. We have taken a total of 438 key-frames in 10 categories from seven different lecture videos that range in length. First, binary and greyscale versions of the input colour images are created. Before using the OCR APIs, the frames are additionally preprocessed to improve the image quality. The recognition accuracy demonstrated that the GCV OCR performs effectively, saving computing time by collecting image text with the highest accuracy of other tools, 96.7 percent.**

Keywords*—Video lectures; keyframes; Google cloud vision (GCV); Tesseract; Abbyy Finereader; Transym; text extraction

I. INTRODUCTION

A branch of machine learning known as optical character recognition (OCR) is focused on identifying characters in visuals such as scanned papers, printed books, or photographs. Despite being a promising technology, there are currently no OCR solutions that can reliably recognise every type of text. Machines can directly handle texts found in the current world thanks to optical character recognition [13]. Education, banking, government, and medical sectors are just a few of the industries where OCR is used. The pre-processed image is fed into the OCR Engine, which then extracts the text that has been written on it. Due to the different written and printed text formats, modern OCR methods use deep learning to increase accuracy. The issue of text recognition can be solved using a variety of conventional deep learning techniques. The most well-known ones include YOLO [1, 2], SSD [2], Mask RCNN [3], and Faster RCNN [2]. These designs may be trained to do character recognition and are essentially entity detectors. Region-based detectors use algorithms like Faster RCNN and Mask RCNN. This implies that the method first scans the image for objects (text) before classifying them (characters). It is slower but more accurate because of this two-step approach. Single Shot Detector (SSD) algorithms like YOLO and SSD simultaneously scan the items and classify them. They are quicker because of the single step procedure, but they do poorly with smaller items, like text in our example. These systems are trained on any of the aforementioned datasets, and the trained systems can be used to anticipate or identify the text in any given image. The goal of qualified neural network (NN) rule generation has spurred a variety of research efforts. The primary classification method for such algorithms is in the manner in which they generate rules. The decomposition method compares each hidden and production node separately, and a pattern is derived from it for precise word detection from images. In a feed-forward NN, each neuron's output is quantified as:

$$R_j = \left(\left(\sum_i W_{ij} \times A_i\right) + \vartheta_j\right) \quad (1)$$

where $A_i = \frac{1}{1+e^{-ax}}$

here, A is the level of activation of neuron i, W_{ij} represents the weight of the relationship between neuron i and j, and is the level of activation of neuron j that controls the gradient of the sigmoid. The breakdown method's most important feature is that almost all of neurons in the NN have either 0 or 1 activations. Binary inputs trigger this in the hidden layer's neurons.

Numerous artificial intelligence scholars have attempted to address the issue of OCR difficulty in order to develop effective OCR systems able to operate in an accurate and timely manner since the advent of computerised systems [28–30]. Even though there are a variety of OCR techniques and toolkits now accessible in the literature, we will be comparing four popular OCR toolkits: Google Cloud Vision (GCV) OCR [24], Tesseract [25], ABBYY FineReader [26], and Transym [27].

Due to the enormous amount of data that deep learning demands for model training, businesses like Google have an advantage in achieving promising outcomes with their OCR services. The specifics of Google Vision OCR are covered in this paper. Using a straightforward REST API interface, the GCV API [7] constructs highly complicated machine learning models focused on image recognition. It has a wide range of image recognition abilities. In this paper, we've concentrated on the OCR module, which scans an image for text before parsing it into data for our computers to use.

A. Objectives

The following are the objectives of this study:

- Data Acquisition by extracting key-frames from lecture videos
- Pre-process the raw input dataset to improve the image quality
- Appy OCR engines to extract text from the key-frames
- Compare the performance OCR engines to decide the best OCR

II. Literature Survey

Deep learning is used in computer vision to build NNs that direct image analysis and evaluation [23]. The OCR methods were mechanical machines, not computers, that could recognise characters at first, but the performance was extremely slow, and the results were less accurate. Although OCR is not a recent issue, its roots can be seen in methods used before the development of computers [12]. OCR has been applied in a wide range of fields. The Transym and Tesseract OCR technologies, for instance, were used by Patel and Patel to analyse car licence plates [17].

The paper [18] used the GCV API to analyse images in another scenario involving an autonomous vehicle to increase the accuracy of object identification and give tough-environment autonomous robots the capacity to recognise objects. Additionally, many industries employ this system to speed up data entry and decrease human error when removing information from document management systems [19], [20]. Additionally, such innovation has been used more and more in smart systems, cloud computing, IoT, and robots. Examples include IoT-based car verification systems [22] and road sign text interpretation [21].

On text in floor plan pictures, conventional and deep learning text detection techniques were contrasted [14]. Four approaches were compared in the study: EAST, Maximally Stable Extremal Regions (MSER), Connectionist Text Proposal Network (CTPN), Stroke Width Transform (SWT), Tesseract, and a normal image processing methodology are the first four options. The last option combines all four of the first three options. Extra sub images were employed for the CTPN approach at the border since CTPN had trouble reading text that was near to the picture borders [14]. The combined technique produces an output that depends on voting by comparing the outcomes from all three previous methods against one another. All approaches to combining particular text boxes into a single text item underwent post processing. Initially, the text was categorised according to the rules. Next, room characteristics were compared to a dictionary of acceptable terms, and the nearest keyword was substituted according to edit distance and term frequency. The proposed approaches were tested on datasets with different levels of quality. The noisy and low quality images were demonstrated to have substantially reduced efficiency with the CTPN approach. The combination technique had the best accuracy on the poor quality images, while the EAST approach seemed to have the greatest recall and F1-score. The efficiency of the detected text was not thoroughly examined, and none of the suggested algorithms could recognise slanted or curving text items. However, it was reported that Tesseract did not make correct estimates on the low resolution pictures.

For image analysis, the GCV API was utilised [8]. Their effort locates and recognises printed text hidden within images, as well as particular items and faces inside images. The adaptability of the GCV API to input noise is assessed in the paper. In particular, when noise is applied to a group of images, the API would be unable to identify the appropriate text or object since, when the noise is cleared, the output is equivalent to the original image. Noise filtering is available for the GCV API. A model that enables users to hear the image's main message in their own language has been proposed in [9]. Text is first taken from the picture and afterwards transformed into the person's native language speech. After being captured by the camera, the image is converted to text by the OCR engine. The gTTS is then used to translate text into speech [9]. A system that interprets words from a taken image has been suggested. Tesseract OCR is used to extract text from digital documents, and the text is then converted to voice. In order to reduce noise, the first acquired image is first transformed to grayscale. After using thresholding, the image is transformed to a binary format, cropped, imported into tesseract OCR for word recognition, and outputted as a text file that can be used as an input for E-speak to generate audio [10]. As of right now, any language's text can be manually entered and converted to any other language as necessary. A whole text book's images cannot be translated from one language to another. Some mobile applications that attempted to convert the above exhibited significant faults. The current system [11], which uses classic OCR, is unable to distinguish text from blurry or poor resolution, blurriness, high noise, and distorted images. The final product is distorted by the image noise. Consequently, consumers experience a challenge with comprehension.

This study primarily focuses on text extraction from lecture slides using Google Cloud Vision (GCV), Tesseract, Abbyy Finereader, and Transym OCR and compares the results to develop a lecture video indexing scheme for the non-linear steering in lecture videos to watch only the interesting points of topics. The dataset is total of 438 key-frames in 10 categories from seven different lecture videos that range in length. First, binary and greyscale versions of the input colour images are created. Before using the OCR APIs, the frames are additionally preprocessed to improve the image quality. The recognition accuracy demonstrated that the GCV OCR performs effectively, saving computing time by collecting image text with the highest accuracy of other tools.

III. Steps Involved in OCR

OCR is a programme that converts text into an appropriate machine-readable format [18, 19]. OCR technology is often used in businesses for automation and processing of written receipts [20]. Researchers now have access to a wide collection of electronic texts that can be analysed using just a few keywords thanks to the OCR technique. Fig. 1 depicts the general OCR approach for text extraction:

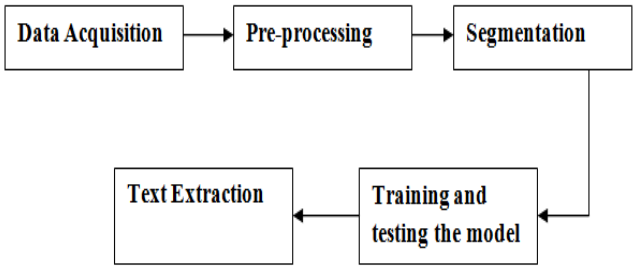


Fig. 1. General Text Extraction using OCR Model.

A. Data Acquisition

The data is a picture of text with straightforward or intricate layouts or backdrops in a scene or document from nature. We can get the text's visual representation via digital camera and handheld scanner [15]. There are several different types of text image databases that can be used for study. They are used to establish standards for processing speed, accuracy, and storage. A few of the datasets available for text extraction is given in [16].

B. Pre-Processing

Before using the OCR method, the raw input dataset must be cleaned up in this step to improve the image quality. The input image must be turned to grayscale and gaussian blur. The 1-dimensional and 2- dimensional Gaussian formula is given below in equations 2 and 3, respectively.

$$\text{GB (i)}=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{i^2}{2\sigma^2}} \quad (2)$$

$$\text{GB (i, j)} = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{i^2+j^2}{2\sigma^2}} \quad (3)$$

where i and j are the horizontal and vertical axis's distance from the origin respectively, and σ is the Gaussian distribution's standard deviation.

C. Segmentation

The pre-processed images are divided into several sections during segmentation. This comprises scanning an image for clusters of pixels that contain character-containing elements; each of these elements has a class applied to it. Any thresholding method must be used in order to allow for additional analysis. In general, using the right settings makes adaptive thresholding operate best. The segmentation procedure is carried out as follows:

$$S_\sigma(I[l_m, l_n]) = \sqrt{\sum_{I=l_i}^{m} prbden(i) + \left(T - l_i\left(I[l_i, l_j]\right)\right)^2 \times p_l^{I[l_i,l_j]}} + \theta \quad (4)$$

where $prbden$ (probability density) is,

$$prbden(i) = \frac{Histogram\,(i)}{mn} + threshold\ + intensity\,(i) \quad (5)$$

for i=0,1,…..n-1

A feature for an immediate layer's adaptive pixel set is calculated as follows:

$$o_i = \mu p_{i(i)} + \delta(i,j) \quad for\ i = 1,2,\ldots\ldots\ldots m \quad (6)$$

$$o_i = \mu q_{i(j)} - 2(x) + \delta(i,j) \quad for\ j = 3,4,\ldots\ldots n \quad (7)$$

here, μ is the layer importance taken into account when organising the layers in a sequential manner for precise and distinctive text recognition. The created single layer has a fixed total and estimates the result as the sum of all the pixels which make up a set. A fresh layer is produced as:

$$o_i = \frac{\sum \overline{w}(i)h(i)+\theta}{\sum \overline{w}(i)} + \delta(j,k)f_i \qquad for\ k = 1,2,\ldots\ldots m-n \quad (8)$$

D. Training and Testing the Model

The crucial OCR phase is model training. Numerous hyperparameters are engaged in this situation. These have either been generated from the training data or have default values set. Following their definition, a model that creates a generic picture -> text modelling for the data processes the data in the training event. The below Fig. 2 depicts the training and testing phase of the model.

On the provided image, feature extraction is carried out using FEL to produce a feature map. CRGL uses a 3×3 hole convolutional and the anchor procedure to create basic areas in an image by clearing dublicate features. CASL uses Soft-Nonmaximum Suppression (NMS) to get the preliminary areas in the image. A Region of Interest (RoI) can be obtained from the image by using the ROI pooling method in TDL. The training model is established as:

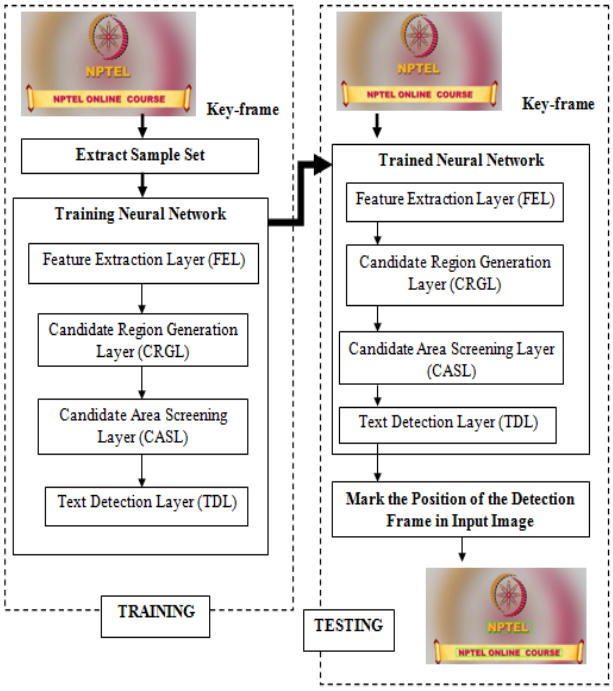


Fig. 2. Training and Testing Phase.

$$TR_1(B,A) = \sum_{i=0}^{m}\sum_{j=0}^{n}\left[\begin{array}{c}\sqrt{\frac{\mu_A^2(a_y)+\mu_B^2(b_y)}{2}} \\ +\sqrt{\frac{1}{m-1}\sum_{i=1}^{n}\left|p_{i-\mu}\right|^2} \\ +\sqrt{\frac{(1-\mu_A(a_y))^2+(1-\mu_B(b_y))^2}{2}}+\end{array}\right] \quad (9)$$

Since NN is used, the hidden layers are taken into account for precise text extraction. Each hidden layer's input is:

$$H(I(i,j)) = I_{i,j} + \sum F_i * W \quad (10)$$

where W and F are the weights between hidden layer and input, and the hidden layer's bias value respectively. As a result, the output of each hidden layer is determined using:

$$o(I(i,j)) = \frac{\sum_{j=1}^{n}(A_{ij}^{M}+B_{ij}^{N})x_j}{\sum_{i=1}^{m}(A_{ij}^{M}+B_{ij}^{N})} \quad (11)$$

E. Text Extraction

To increase the model's ability to extract text accurately, an analysis step is taken after processing through first four steps. The text that was retrieved from the image is given by the following pixels:

$$T(X,Y) = \sum_{(i,j)\in F_s, i\in W, j\in T} o(i) + W_i + \frac{(i*j)(i,j)}{(H_{i,j(P,Q)})} \quad (12)$$

An interface utilising Google OCR technology has been developed in order to provide users with a simple and practical method of text extraction from images. Additionally, this will automate a few processes through the use of the Google OCR engine. The goal of this work is to extract text from English-language lecture slides. The user can choose a language and start the text extraction process. For the purposes of OCR, the regions are separated into unoccupied and occupied regions. Following that, a machine learning model is used to scan the data before a number of processes, including area segmentation and extraction, creating the necessary line images for line segmentation inputs, ground truth output, and more.

IV. Proposed OCR in Slide-to-Text (STT) Conversion

With an emphasis on image recognition, the GCV API transforms extremely complicated machine learning models into a straightforward REST API interface. We concentrate on the OCR module in this work. A Python script was used to construct the Fig. 3 workflow in Tensorflow.

- We have considered seven different lecture videos (machine learning, network, DBMS, Algorithms, two cryptography, and data science for engineers) of variying duration.
- We have first extracted the key-frames (images) from each lecture videos [total 438 images of 10 categories, including: 1) Digital Images, 2) Machine-written characters, 3) Hand-written characters, 4) Machine-written digits, 5) Hand-written digits, 6) Multi-oriented text strings, 7) Black and white images, 8) Noisy images, 9) Skewed images, and 10) Blurred images].
- The obtained images are transformed from the colour to grayscale and binary images. The pre-processing procedures (sharpening, contrast adjustment, and brightness adjustment) are also used to improve the image quality before applying the OCR APIs.
- The processed images are then uploaded to Google Cloud Storage (GCS). Vision API and background processes are started by a GCS event to Create a transcription of the GCS-stored image.
- The converted images are yet again saved in GCS for use in the future. The Natural Language API is used to extract entities from the converted images. The tool initially segments the image's structure to determine where the text is located. The OCR module then does a text recognition on the proper area to generate the text after detecting the general location.
- In a post-processing step, errors are finally fixed by running the data through a language model. The convolutional neural network (CNN) used to do all of this merely connects each neuron to a portion of the neurons in each layer. CNN is designed to mimic the hierarchical organisation of our visual system in terms of object (characters) recognition.

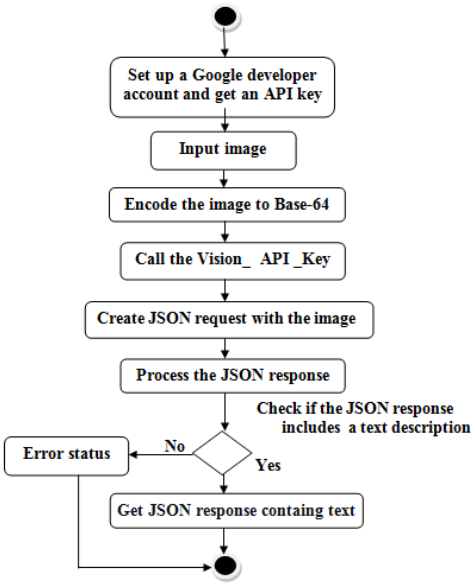


Fig. 3. Text Extraction using Google OCR.

V. Results and Discussion

This paper used a desktop computer with an i8 processor, 8 GB of RAM, 512 GB of storage, and an HDD (Hard Disk Drive). The text extraction results from each lecture video using the acquired key-frames demonstrated that GCV performed better than other OCR APIs in extracting text from the key-frames, with an average accuracy of 96.7 percent, as shown in Table I. Tesseract's is 92 percent, Abbyy Finereader's is 90.5 percent, and Transym's is 80.8 percent.

TABLE I. A COMPARISON OF OCR APIS

Dataset (key-frames of LV)	Method	Pr (%)	Re (%)	F1-Score (%)
38	Google OCR	97.2	94.7	**97.4**
	Tesseract	88.2	89.4	88.7
	Abbyy Finereader	87.8	86.8	87.2
	Transym	65.6	84.2	73.7
39	Google OCR	94.7	97.4	**96.0**
	Tesseract	86.1	92.3	89.0
	Abbyy Finereader	91.4	89.7	90.5
	Transym	72.7	84.6	78.1
38	Google OCR	97.2	97.3	**97.2**
	Tesseract	88.8	94.7	91.6
	Abbyy Finereader	88.2	89.4	88.7
	Transym	62.5	84.2	71.7
75	Google OCR	97.2	97.3	**97.2**
	Tesseract	91.5	94.6	93.0
	Abbyy Finereader	91.3	92.0	91.6
	Transym	83.3	88.0	85.5
72	Google OCR	98.5	98.6	**98.5**
	Tesseract	92.6	94.4	93.4
	Abbyy Finereader	95.5	93.0	94.2
	Transym	86.1	90.2	88.1
51	Google OCR	98.3	96.0	**97.1**
	Tesseract	93.4	90.1	91.7
	Abbyy Finereader	88.6	86.2	87.3
	Transym	75.6	80.3	77.8
125	Google OCR	89.4	98.4	**93.6**
	Tesseract	96.6	96.8	96.6
	Abbyy Finereader	94.1	95.2	94.6
	Transym	88.8	93.6	91.1

The GCV OCR's accuracy is much higher than that of other techniques while taking into account the file size and resolution. Additionally, the accuracy of the low-resolution or small-size images is the lowest. The three parameters listed below are used to evaluate performance.

$$Recall\ (Re) = \frac{Extracted\ text}{Total\ key-frames} \quad (13)$$

$$Precision(Pr) = \frac{Correctly\ extracted\ text}{Extracted\ text} \quad (14)$$

$$F1 - score = \frac{2 \times Re \times Pr}{Re + Pr} \quad (15)$$

The images were reduced to 720 x 480 pixels because the more pixels an image has, the longer OCR would take to process it into grayscale. To cut down on the amount of time needed for STT translation, all preprocessing stages were completed. Everything in the GCV OCR is contained within a RESTful API that provides a JSON structure with the text and bounding box (containing image text area with x and y co-ordinates). It takes about 15 seconds to translate a STT. The sample output of text extraction using GCV OCR is shown in Fig. 4. Precision, recall, and F-score of different OCR APIs is shown in Fig. 5.

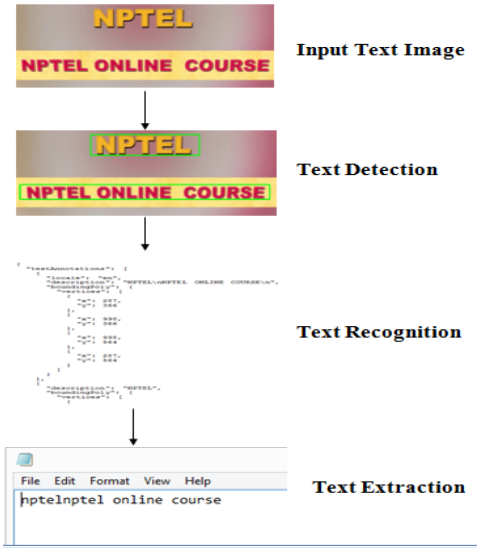


Fig. 4. Text Extraction from an Image.

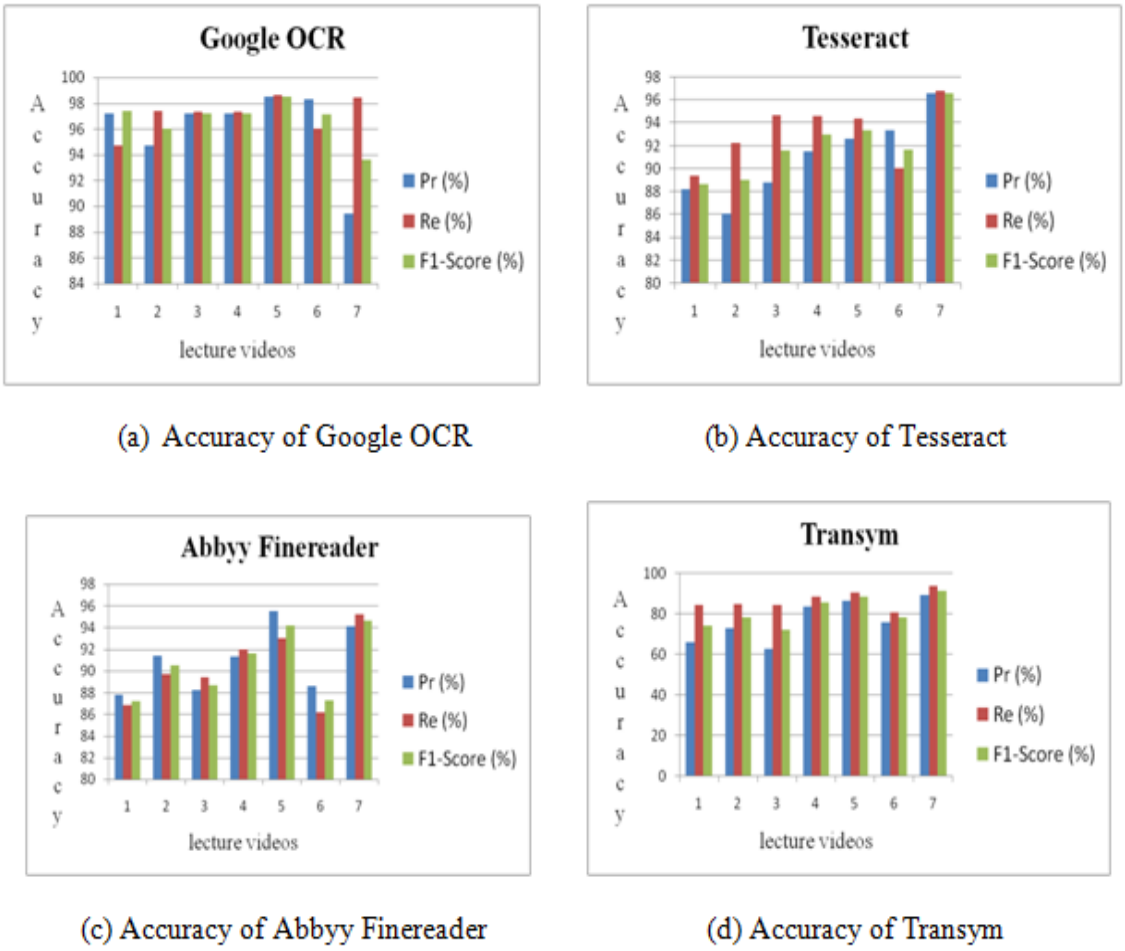


Fig. 5. Precision, Recall, and F-Score of Different OCR APIs.

From this result we can clearly say that the GCV OCR is much better than Tesseract, Abbyy Finereader, and Transym, with accuracies of 96.7%, 92.0%, 90.5%, and 80.8%, respectively, in STT conversion (shown in Fig. 6).

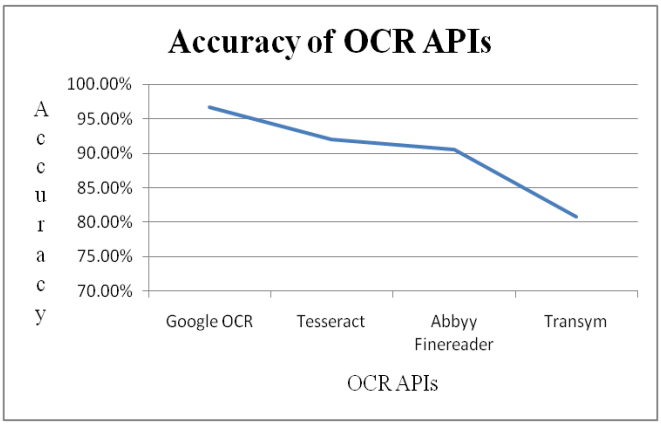


Fig. 6. Performance Comparison of Different OCR APIs.

A comparison of a number of quality criteria provided by the OCR systems is summarised in Table II.

A. Discussion

In order to make the tools more effective in identifying and processing information, this section addresses some noteworthy results, fascinating difficulties, and other usage domains or areas of study. In terms of size and image attributes, the GCV API is more accurate than competing APIs. In terms of additional factors, we discovered the following:

- All the tools were able to recognise English letters with comparable proficiency.
- Slightly elevated images could be detected by all the tools with a high degree of accuracy, while very small, distant, or blurry images could not be recognised by both the Abbyy Finereader and Transym tools.
- The supplied image's watermark background and grey-colored text significantly lower the text identification performance.

TABLE II. OCR SYSTEMS

OCR methods	Pros	Cons
Transym [27]	• Available as a SDK • Multilingual support • Support machine written characters	• Not available online • Not open source
ABBYY Finereader [6]	• Best for business users • Supports Automation • Batch processing • Support for 192 languages	• Not for general users • Not open access
Tesseract [4] [5][6]	• Quite powerful and accurate • Supports over 100 languages	• Not for business users
GCV API [4]	• Quick and easy OCR software for general users • Support for over 200 languages • Mobile app support • Available on almost all platforms • Quite powerful and accurate	• Not open access

The Tesseract and GCV APIs outperform the other two. Because Tesseract is open-source software that can be developed, customised, and managed according to particular needs, it is great software for developers. Tesseract, however, can be somewhat challenging to install and configure. Due to the availability of a variety of services, the GCV API performs better than Tesseract. It is also straightforward to connect to, configure, and use services on. The following are some potential strategies to enhance the functionality of OCR technologies to make them more effective at recognising and evaluating information:

- To cut down on extra reading material and prevent wrongly positioned images, the programmer should define the border, frame, or template matching.
- Before the recognition process, the programmer should make any necessary colour adjustments to the character, as well as remove the extra watermark backdrop.
- To aid with the understanding difficulties with the presentation slides, the programmer should create a programme that can connect models, enabling both printed and handwritten text recognition.
- The effectiveness of the post-processing outcomes can be increased by using natural language processing techniques.

VI. CONCLUSION

Extracting text from lecture slides is crucial for indexing the lecture video. This study evaluates the text extraction capabilities of the GCV OCR, Tesseract, Abbyy Finereader, and Transym in order to develop a lecture video indexing scheme for the non-linear steering in lecture videos so that viewers only watch the interesting points of topics. According to the test findings, Google Cloud Vision had accuracy rates of 96.7 percent, 92.0 percent, 90.5 percent, and 80.8 percent, which were higher than those of Tesseract, Abbyy Finereader, and Transym. The amount of time needed for processing an image grows as its resolution does. In order to reduce the time needed for STT translation, the images are first reduced to 740 x 480 pixels and then converted to grayscale. According to this study, resizing and preprocessing an image before performing OCR can greatly increase the OCR's accuracy. It takes about 15 seconds to translate an STT. This study gives an idea for the researchers who work on OCR.

Our future work will include an effort to assess additional OCR services utilising substantial datasets and more statistically significant analyses for their accuracy and durability. We will make use of cutting-edge image processing techniques and assess how they may be used to create OCR systems that are more precise and effective. In the future, we'll also work on turning the audio from the lecture into text and creating the index points using an effective ASR tool. The results of this study will help to generate index points.

ACKNOWLEDGMENT

We are very thankful to our parents, family, and friends for supporting to complete this work.

REFERENCES

[1] Shashidhar R, A S Manjunath, Santhosh kumar R, Roopa M, Puneeth S B. “Vehicle Number Plate Detection and Recognition using YOLO- V3 and OCR Method” 2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC) . pp. 1-6 (2021). https://doi.org/10.1109/ICMNWC52512.2021.9688407.

[2] Shrey Srivastava, Amit Vishvas Divekar, Chandu Anilkumar, Ishika Naik, Ved Kulkarni and V. Pattabiraman. “Comparative analysis of deep learning image detection algorithms.” Journal of big data. Pp. 1-27 (2021). https://doi.org/10.1186/s40537-021-00434-w.

[3] Canhui Xu, Cao Shi , Hengyue Bi , Chuanqi Liu, Yongfeng Yuan, Haoyan Guo, And Yinong Chen. “A Page Object Detection Method Based on Mask R-CNN.” IEEE Access. Pp. 143448- 143456 (2021). https://doi.org/10.1109/ACCESS.2021.3121152.

[4] Thomas Hegghammer. “OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment.” Journal of Computational Social Science (2022) 5:861–882. https://doi.org/10.1007/s42001-021-00149-1.

[5] Malathi T , Selvamuthukumaran D , Diwaan Chandar C S , Niranjan V , Swashthika A K. “An Experimental Performance Analysis on Robotics Process Automation (RPA) With Open Source OCR Engines: Microsoft Ocr And Google Tesseract OCR.” IOP Conf. Series: Materials Science and Engineering. Pp. 1-9 (2021). doi:10.1088/1757-899X/1059/1/012004.

[6] Abdulkarim Malkadi, Mohammad Alahmadi, Sonia Haiduc. “A Study on the Accuracy of OCR Engines for Source Code Transcription from Programming Screencasts.” 2020 Association for Computing Machinery. ACM. https://doi.org/10.1145/3379597.3387468.

[7] https://research.aimultiple.com/ocr-accuracy/

[8] Hossein Hosseini, Baicen Xiao and Radha Poovendran “Google’s Cloud Vision API Is Not Robust to Noise,” 16th IEEE International Conference on Machine Learning and Applications December 18-21, 2017.

[9] Rithika.H, B. Nithya santhoshi “Image Text to Speech Conversion in The Desired Language by Translating with Raspberry Pi, ” International Conference on Computational Intelligence and Computing Research 2016.

[10] Yasuhisa Fujii, "Optical Character Recognition Research at Google", IEEE 7th Global Conference on Consumer Electronics (GCCE), December 2018.

[11] Mr. Rajesh M., Ms. Bindhu K. Rajan Ajay Roy, Almaria Thomas K, Ancy Thomas, Bincy Tharakan T, Dinesh C “Text recognition and face detection aid for visually impaired person using raspberry pi” International Conference on circuits Power and Computing Technologies [ICCPCT] July 2017.

[12] Mr. Rishabh Dubey "Machine Learning in the Field of Optical Character Recognition (OCR)" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-5, August 2020, pp.1664-1668, URL: https://www.ijtsrd.com/papers/ijtsrd33233.pdf

[13] Devices Yu Weng and Chunlei Xia. “A New Deep Learning-Based Handwritten Character Recognition System on Mobile Computing.” *Mobile Networks and Applications* volume 25, pages402–411 (2020). https://doi.org/10.1007/s11036-019-01243-5.

[14] J. Ravagli, Z. Ziran, and S. Marinai, “Text recognition and classification in floor plan images,” in 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 1, Sep. 2019, pp. 1–6.

[15] Liang, J., Doermann, D. and Li, H. (2015) “Camerabased analysis of text and documents: a survey”, International Journal on Document Analysis and Recognition, pp. 1-21.

[16] Lingqian Yang, Daji Ergu, Ying Cai, Fangyao Liu, Bo Ma. “A review of natural scene text detection methods.” The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021). Procedia Computer Science 199 (2022) 1458–1465. https://doi.org/10.1016/j.procs.2022.01.185

[17] C. Patel, A. Patel, and D. Patel, “Optical character recognition by open source OCR tool tesseract: a case study,” International Journal of Computer Applications, vol. 55, no. 10, pp. 50–56, Oct. 2012, doi: 10.5120/8794-2784.

[18] M. Sugadev, Yogesh, P. K. Sanghamreddy, and S. K. Samineni, “Rough terrain autonomous vehicle control using Google Cloud Vision API,” in 2019 2nd International Conference on Power and Embedded Drive Control (ICPEDC), Aug. 2019, pp. 244–248, doi: 10.1109/ICPEDC47771.2019.9036621.

[19] J. Sharma, G. S. Sindhu, S. Sejwal, J. Solanki, and R. Majumdar, “Intelligent vehicle registration certificate,” in 2019 Amity International Conference on Artificial Intelligence (AICAI), Feb. 2019, pp. 418–423, doi: 10.1109/AICAI.2019.8701286.

[20] F. Adamo, F. Attivissimo, A. Di Nisio, and M. Spadavecchia, “An automatic document processing system for medical data extraction,” Measurement, vol. 61, pp. 88–99, Feb. 2015, doi: 10.1016/j.measurement.2014.10.032.

[21] I. Kavati, G. K. Kumar, S. Kesagani, and K. S. Rao, “Signboard text translator: a guide to tourist,” International Journal of Electrical and Computer Engineering (IJECE), vol. 7, no. 5, pp. 2496–2501, Oct. 2017, doi: 10.11591/ijece.v7i5.pp2496-2501.

[22] A. J. Samuel and S. Sebastian, “An algorithm for IoT based vehicle verification system using RFID,” International Journal of Electrical and Computer Engineering (IJECE), vol. 9, no. 5, pp. 3751–3758, Oct. 2019, doi: 10.11591/ijece.v9i5.pp3751-3758.

[23] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, “Computer vision and deep learning techniques for pedestrian detection and tracking: A survey,” Neurocomputing, vol. 300, pp. 17–33, Jul. 2018, doi: 10.1016/j.neucom.2018.01.092.

[24] Google drive (2012). http://drive.google.com

[25] S mith, R.: An overview of the tesseract OCR engine (2007)

[26] Abbyy OCR (2016). https://www.abbyy.com/

[27] Transym OCR. http:// www.transym.com/download.htm

[28] Patil, V.V., Sanap, R.V., Kharate, R.B. “Optical character recognition using artificial neural network.” Int. J. Eng. Res. Gen. Sci. 3(1), 7 (2015)

[29] Bautista, C.M., Dy, C.A., Ma˜nalac, M.I., Orbe, R.A., Cordel, M. “Convolutional neural network for vehicle detection in low resolution traffic videos.” In: 2016 IEEE Region 10 Symposium (TENSYMP), pp. 277–281. IEEE (2016)

[30] Ye, Q., Doermann, D. “Text detection and recognition in imagery: a survey.” IEEE Trans. Pattern Anal. Mach. Intell. 37(7), 1480–1500 (2015).

Disease Prediction Model based on Neural Network ARIMA Algorithm

Kedong Li*
Jining Center for Disease Control and Prevention
Jining, China

***Abstract*—Because the morbidity data of infectious diseases do not only have a single linear or nonlinear characteristic, but also have both linear and nonlinear characteristics, the combination model prediction method is often used to predict the morbidity of infectious diseases in recent years. Compared with the single model prediction analysis method, the combination model can combine the advantages of a single model to extract the effective information contained in the original time series more scientifically and fully. In the context of big data, for the medical field, massive medical data is complex, and the traditional manual data processing method has been unable to meet the current needs. With the help of the computer, data mining can discover new knowledge that is potentially useful and understandable by clearing, integrating, selecting, and transforming the original data. Using data mining, we can organize and reproduce the useful medical knowledge hidden in medical big data. In this paper, an ARIMA-GRNN model is established; the fitting value and the corresponding time are used as the input of the neural network. The actual morbidity is used as the output to train the network and construct the ARIMA-GRNN combined model. Due to the different information flow of BP neural network and neural network, this study also constructed ARIMA-GRNN combined model and ARIMA model, and compared the modeling effect and prediction performance of various models. The average absolute percentage error of the experimental results in this paper is less than 8.63%, and the average absolute percentage error is less than 5%. Compared with other models, it has a better prediction effect, higher accuracy, and more obvious advantages. In this paper, the prediction of disease is dynamic and continuous. It is of great significance for disease prevention and control to use monitoring data to study the epidemic trend and periodic change law, and to make a reasonable prediction.**

Keywords*—Disease prevention and control; trend prediction; neural network; combination model; ARIMA algorithm

I. INTRODUCTION

At present, feature learning technology is mainly divided into two categories: domain knowledge-driven and data-driven. Domain knowledge- driven methods extract features from image data or non-image data based on domain knowledge. For image data, it is mainly to extract low-level features such as artificially designed texture and wavelet, or qualitative indicators such as capsule and peritumoral blood vessels [1]. Non-image data mainly include some clinical characterization indicators of liver tumors, such as quantitative indicators, age, and gender, or qualitative and laboratory indicators, such as protein level and liver function, quantitative and qualitative characterization indicators in non-image data, and artificial design features in image data. The heterogeneity between the two brings difficulties to knowledge-based feature learning. From another point of view, the representation fusion between image data and non-image data brings hope to the performance improvement of the feature learning model [2]. Knowledge-driven feature learning method has good interpretability and robustness through artificial feature extraction and algorithm design, but its disadvantage is that it needs artificial participation, low efficiency, and difficult model processing. Data-driven methods are represented by deep learning techniques.

The data-driven neural network has a powerful self-learning ability, which can automatically learn features from a large number of data samples, and is the mainstream of data-driven methods [3]. With the increasing number of convolutional neural network layers and the increasing network width, the low-level pixel-level features can be gradually abstracted into high-level semantic features layer by layer to better extract the rich features hidden in large-scale image data [4]. As an effective learning method for extracting high-level semantic features, convolutional neural networks have achieved good results in many image classification and segmentation tasks. Data-driven feature learning methods are more efficient, but less interpretable and robust, and depend heavily on the number of labeled samples [5]. Especially when the convolutional neural network which has achieved better performance in the field of natural images is directly transferred to the field of medical images with small samples. The network is easy to overfit and has poor robustness [6]. Thus, in small-scale medical image feature learning, the mainstream CNN network is still difficult to meet the VMI. The structure of the network, the training mode, and the scale of the parameters need to be redesigned and adjusted to suit the specific VMI prediction task.

In 2015, Wei Wu et al. used the ARIMA model, ARIMA-GRNN combined model, ARIMA, and feedback dynamic nonlinear autoregressive neural network to establish a combined model based on monthly incidence data of hemorrhagic fever with renal syndrome. The study showed that the prediction performance of the combined model based on a dynamic neural network was higher than that of the static neural network. The prediction accuracy of the combination model is higher than that of the single model [7]. In 2016, Tian Dehong used the monthly incidence data of human brucellosis in China to establish the ARIMA model, BP neural network, ARIMA, and BPNN to establish the combination model. The study showed that the prediction accuracy of the combination

*Corresponding Author.

model was significantly higher than that of the single model [8]. In 2017, Wang Yongbin and others used the incidence data of hand, foot and mouth disease to establish ARIMA model, RBF neural network and ARIMA and RBF neural network to establish a combination model. The results showed that the prediction accuracy of the combination model was better than that of the RBF neural network model and ARIMA model [9]. The above literature analysis shows that compared with the single model prediction analysis method, the combination model can combine the advantages of a single model to extract the effective information contained in the original time series more scientifically and fully.

In this study, the ARIMA model was chosen to establish neural network and Elman neural network, considering that the predicted value of ARIMA model can fit the seasonality and periodicity, and has a similar trend to the measured value. Therefore, the fitting value and time of the disease Arima model are used as the input of the network. The actual morbidity is used as the output to train the network. Its core is to use the nonlinear mapping ability of the neural network to correct the random effect part to improve its prediction accuracy. The realization of the prediction model can realize the prediction of the disease epidemic situation and assist the medical staff to predict and manage the disease epidemic situation. Therefore, this study can effectively assist in the diagnosis and prognosis of the disease, and play an unlimited clinical and social value on the basis of the use of limited medical resources, especially in critically ill patients.

The main contents of this paper are as follows:

1) The background and significance of the research are introduced.

2) The basic theory of the prediction model is introduced.

3) The modeling steps of the BP neural network are analyzed.

4) The empirical part of the combination model is done.

5) Comparison of the fitting diagram of the combined model, comparison of the actual prediction effect and conclusion are done.

6) Conclusions and prospects for the whole paper are made.

II. Related Work

A. Artificial Neural Network

ANN (Artificial Neural Network) is an artificial model based on the function of the human brain and connected with various problems in real life with the relationship between mathematics and physics [10]. The research of artificial neural networks is based on the structure of the biological nervous system. The smallest element in the nervous system of the biological world is the neuron, which consists of nerve cells and multiple processes [11]. The artificial neural network is constructed with reference to biological neurons, and the composition of a single artificial neuron is shown in Fig. 1.

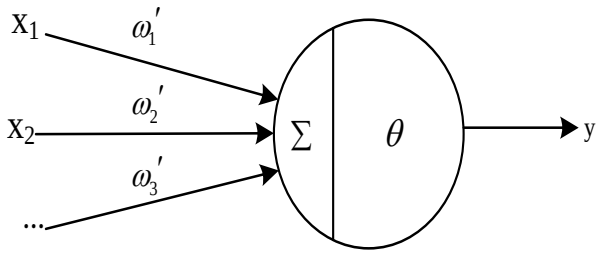


Fig. 1. Diagram of Single Neuron Model.

The composition of artificial neurons mainly includes three elements:

1) A series of input signals and the weight at each connection point represent the strength of each signal. When the connection weight is positive, the neuron is activated; and when the connection weight is negative, the neuron is inhibited.

2) A summation module for integrating all the input signals was used [12].

3) A nonlinear activation function, which acts as a nonlinear mapping by limiting the output interval of the neuron is adopted.

In addition, there is a deviation, namely the threshold θ.

All of the above processes are expressed mathematically as follows:

$$net = \sum_{i=1}^{n} X_i \omega_i ; o = f(net - \theta) \tag{1}$$

Where *net* represents the cumulative sum of the input neurons; *o* represents the sum of the neuron outputs; X_i represents the input quantity of the *ith* input neuron, and ω_i represents the connection weight of the *ith* input neuron of this neuron; $f(x)$ is the activation function, which describes the connection between the neuron input and output [13]. The selection of these parameters depends on the size of the training data, the characteristics of the studied sequence and some subjective experience. The quality of parameter selection will play a key role in the final prediction results.

B. Types of Neural Network Structures

The activation functions in the neural network structure are as follows:

1) Hard limit function: The expression for the hard limit function is as follows:

$$y = f(u) = \begin{cases} 1, u \geq 0 \\ 0, u < 0 \end{cases} \tag{2}$$

Or:

$$y = f(u) = \text{sgn}(u) = \begin{cases} 1, u \geq 0 \\ -1, u < 0 \end{cases} \tag{3}$$

Where $\mathrm{sgn}(\square)$ is the sign function. The hard limit function in Equation 2 is also called the single limit function, and the hard limit function in Equation 3 is also called the double limit function.

2) Linear function: The expression for the linear function is as follows:

$$y = f(u) = u \tag{4}$$

The output neurons of neural networks with linear functions realize the function approximation [14].

3) Saturation linear function: The saturation linear function is expressed as follows:

$$y = f(u) = \frac{1}{2}(|u+1| - |u-1|) \tag{5}$$

The curve of the saturation linear function is shown in Fig. 2. This activation function is also commonly used in classification problems.

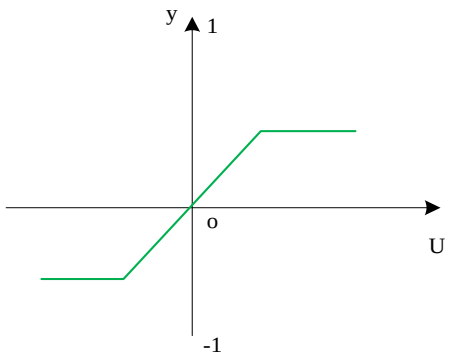


Fig. 2. Saturation Linear Function Diagram.

4) Sigmoid function: Sigmoid function is the most frequently used activation function in neural network algorithms. The Sigmoid function is a strictly monotonically increasing smooth function and has the asymptotic property [15]. The logarithmic tangent function is a form of the Sigmoid function, and its functional form is:

$$y = f(u) = \frac{1}{1+e^{-\lambda u}} \tag{6}$$

In the equation, the parameter λ is called the gain of Sigmoid function, which is the slope parameter of sigmoid function. By changing this parameter, sigmoid functions with different slopes can be obtained. The larger the value of λ, the steeper the curve [16]. The logarithmic tangent function, also known as the unipolar Sigmoid function, is differentiable and varies continuously from 0 to 1 [17]. The plot of the unipolar Sigmoid function is shown in Fig. 3.

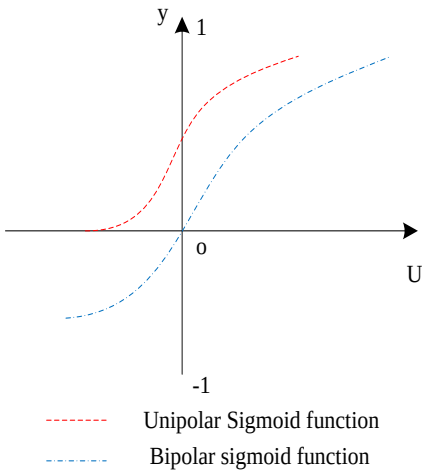


Fig. 3. Sigmoid Function Diagram.

The sigmoid activation function defined in the equation has a range of 0 to 1. Sometimes, the range of the activation function needs to vary from -1 to 1 and be odd symmetric about the origin. A double tangent sigmoid activation function can be used for this purpose [18]. Its functional form is:

$$y = f(u) = \tanh(\lambda u) = \frac{e^{\lambda u} - e^{-\lambda u}}{e^{\lambda u} + e^{-\lambda u}} \tag{7}$$

C. BP Neural Network

In the BP neural network, backward propagation is a learning method requiring supervised learning, which is mainly reflected in the training process of BP neural network. Feedforward network is a structure, which is reflected in the network architecture of BP neural network. A typical feedforward neural network is shown in Fig. 4.

The BP neural network has the characteristics of simple structure, easy to use and high efficiency, which is why more and more studies use the neural network [19]. The algorithm of error backward propagation gradually optimizes the connection weights between neurons by iterative processing so that the error between the final output result and the expected result tends to be stable and minimum.

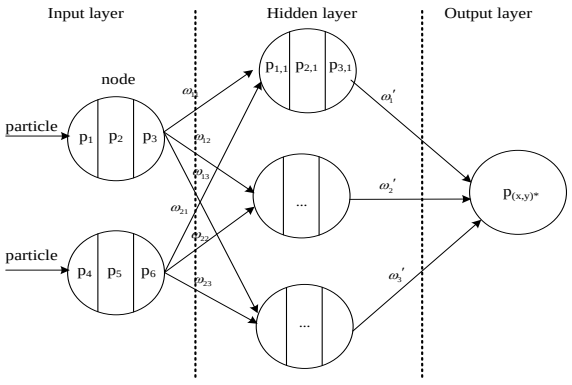


Fig. 4. Typical Feedforward Neural Network Diagram.

III. Modeling Steps of BP Neural Network

The steps of BP neural network modeling and prediction are as follows:

1) Select samples and construct a training set. The use of appropriate samples is an important prerequisite for model construction [20]. Select the appropriate structure according to the actual situation, and try to make the selected structure contain the maximum information.

2) Data preprocessing: the BP neural network has special requirements for the data of training samples. If the data interval changes too much, it cannot be used as output data.

3) Network structure design: this process includes the selection of the number of network layers, the number of hidden layer nodes, the number of input layer nodes and output layer nodes, learning rate, training function, the selection of hidden layer activation function and the output layer activation function.

4) Initialize that network, and randomly distribute the weight value and the threshold value of each connection.

5) Input the divided data.

6) Recalculate and adjust that weight value and the threshold value of each connection in accord to the error.

7) Obtain the latest parameters before proceeding from step.

8) At the beginning, when the given training times are reached or the output error is not higher than the given error standard, terminate the training.

9) Predict that time series by using the model and obtaining a prediction result.

The general process is shown in Fig. 5.

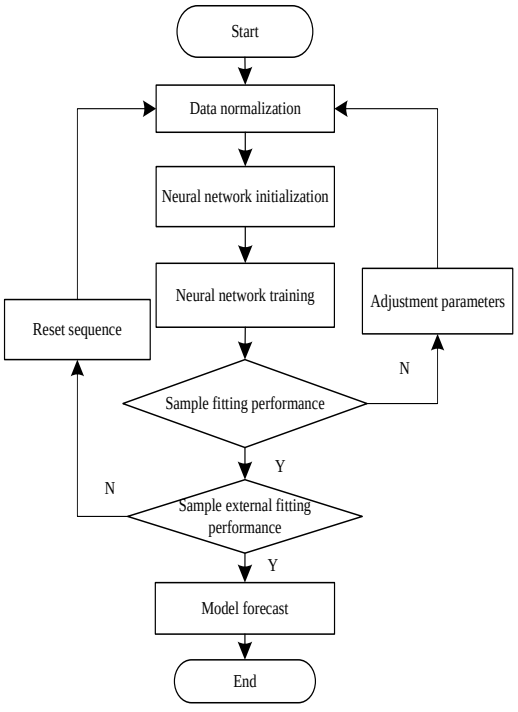


Fig. 5. FLOW Chart of BPNN Model.

IV. Empirical Analysis of BP Neural Network

The raw data is split into two data sets: a training set and a test set. The training set is used to train the model and select the optimal network model; the test set is used to evaluate the performance of the selected optimal network model [21]. The data before January 2021 is used as the training set, and the data for the whole year of 2021 is used as the test set. By dividing the data set in this way, most of the information contained in the original data can be retained so that the network model can be better learned and trained [22]. At the same time, the problems of over-learning and over-fitting can be avoided.

1) Activation Function: in this study, the Sigmoid function is selected as the activation function of the hidden layer, which can well increase the nonlinear mapping ability of the network [23]. Although this study belongs to the regression algorithm, the Sigmoid function is also selected as the activation function of the output layer because the monthly incidence rate of major diseases is between 0 and 1.

2) The number of neural nodes in each layer, the number of iterations and the learning rate: the number of neural nodes in the input layer and output layer is usually determined by referring to the characteristics of their own data. In this study, according to the splitting of the data set, the number of neural nodes in the input layer is 3, and the number of neural nodes in the output layer is 1 [24]. The empirical formula used in this study to estimate the number of neural nodes in the hidden layer is as follows:

$$m = \sqrt{M + N} + a \tag{8}$$

Where, m represents the number of neural nodes in the hidden layer; M represents the number of neural nodes in the input layer; N represents the number in the output layer; and *a* is an adjustment constant ranging from 1 to 10 [25].

Since the amount of data in this study is not too large, the number of iterations is fixed at 1000. The learning rate is 0.15, and the hidden layer is 3-12 for comparison. The data set is shown in Table I.

TABLE I. Training Set MSE and Test Set MSE with Different Number of Hidden Layers

Number	Training set MSE	Test set MSE
3	0.0026	0.0038
4	0.0026	0.0032
5	0.0026	0.0032
6	0.0027	0.0032
7	0.0026	0.0033
8	0.0026	0.0032
9	0.0026	0.0031
10	0.0026	0.0030
11	0.0026	0.0031
12	0.0027	0.0031

In Table I, the selection of the number of hidden layers from 3 to 12 has no obvious effect on the MSE of the training set of data. When the number of hidden layers is from 7 to 11, the MSE of the training set is relatively small. When the number of layers is 10, the test set MSE is the smallest, and the MSE at this time is 0.0030.

V. Experimental Analysis

A. Model Evaluation Comparison Index

Three error evaluation indexes are used to evaluate and compare the prediction effect of each prediction model.

1) Mean Square Error (MSE): The mean square error is the square of the difference between the true value and the predicted value and then averaged over the range [0, + ∞), which is equal to 0 when the predicted value exactly matches the true value, that is, the perfect model; the larger the error, the larger the value. In this study, it refers to the average value of the square sum of the error between the real value and the predicted value of the monthly incidence rate of major diseases. Its calculation formula is formally close to the variance. The formula expression is:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{9}$$

2) Mean absolute error: The mean absolute error is the average of the absolute errors. The absolute error represents the absolute value of the deviation between all observed values and the true value. Its range is [0, + ∞). When the predicted value is completely consistent with the true value, it is equal to 0, that is, the perfect model. The larger the error, the larger is the value. Relative to the average error, because the deviation of the average absolute error is absolute, there is no problem of positive and negative offset between the errors, which can better show the true level of the predicted value error than the average error. In this study, the mean absolute error refers to the mean value of the sum of the absolute value of the deviation between the true value and the predicted value of monthly incidence of major diseases, and its formula expression is:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - \hat{y}_i\right| \tag{10}$$

3) Mean absolute percentage error: The mean absolute percentage error is the absolute percentage deviation of all individual observations from the true value, and its value range is [0, + ∞). A MAPE of 0% indicates that the model is a perfect model, and a MAPE greater than 100% indicates that the model is an inferior model. The mean absolute percentage error has one more denominator than the mean absolute error. The specific expression is as follows:

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{11}$$

Mean square error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) are all affected by the difference between the real value and the predicted value, but all the deviations are squared or absolute in the calculation process, so there will be no negative number, and there will be no positive and negative offset of errors. It can more accurately reflect the error between the predicted value and the real value.

B. Fitting Effect of each Prediction Model on Monthly Morbidity of Major Diseases

The fitting effect between the monthly incidence rate of major diseases obtained by each prediction model and the national monthly incidence rate of major diseases is shown in Fig. 6. Since the fitting data obtained by these prediction models are not much different from the real data, it is not convenient to display them in the same figure, so each model is compared with the real value separately.

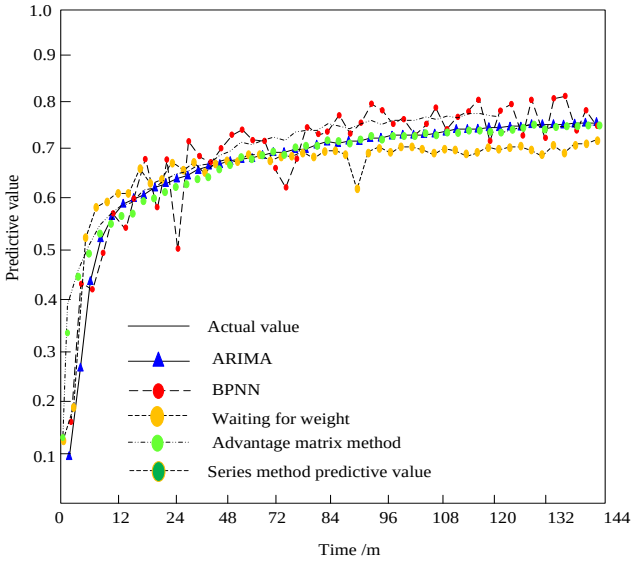


Fig. 6. Comparison between Predicted Value and Real Value by Series Method.

In Fig. 6, the fitting effect of each model is relatively good, which can reflect the change trend of the true value. The fitting effect of the series combination forecasting model is better, especially in the later period.

C. Comparison of Prediction Effect among Different Prediction Models

The annual morbidity data of major diseases in 2021 predicted by all the models used in this study are put into the same table with the real data, as shown in Table II.

TABLE II. PREDICTED VALUES OF EACH MODEL

Time	Real value	BPNN	ARIMA	Equal weight method	Dominance matrix method	Series connection method
2021.1	0.2380	0.2513	0.2367	0.2440	0.2415	0.2273
2021.2	0.1840	0.2483	0.2696	0.2589	0.2625	0.2596
2021.3	0.3835	0.3801	0.4037	0.3919	0.3958	0.4121
2021.4	0.3339	0.3528	0.3471	0.3500	0.3490	0.3446
2021.5	0.4023	0.3731	0.3944	0.3838	0.3873	0.4008
2021.6	0.4178	0.4216	0.4458	0.4337	0.4377	0.4632
2021.7	0.3804	0.3657	0.4078	0.3868	0.3938	0.4170
2021.8	0.4136	0.3778	0.4119	0.3949	0.4006	0.4221
2021.9	0.4427	0.3886	0.4138	0.4012	0.4054	0.4243
2021.10	0.4188	0.3454	0.3614	0.3534	0.3560	0.3613
2021.11	0.5483	0.4343	0.4613	0.4478	0.4523	0.4819
2021.12	0.5680	0.4798	0.4992	0.4895	0.4927	0.5272

By comparing the monthly prediction results of each model with the real data, it can be seen that the prediction effect of the single ARIMA and series combination model is better than that of other models. The advantage of the single ARIMA model in the first few months is more obvious, which is similar to the real value. In the last few months, the prediction effect of the series combination model is better than that of other models. But on the whole, the prediction effect of all models in the last three months from October is obviously not as good as that of the previous months.

The overall evaluation indicators of each model are shown in Table III.

TABLE III. EVALUATION INDEXES OF EACH MODEL

	MSE	MAE	MAPE(%)
ARIMA	0.0021	0.0356	10.0636
BPNN	0.0030	0.0427	11.1487
Equal weight method	0.0024	0.0375	10.1644
Dominance matrix method	0.0023	0.0368	10.0996
Series connection method	0.0016	0.0334	9.6914

The MSE, MAE and MAPE of the series combination model are the smallest. The MSE, MAE and MAPE of the single BP neural network are the largest. Therefore, on the whole, the series combination model shows its advantages.

The average absolute percentage error of the single ARIMA model was 10.06%. The average absolute percentage error of the single BP neural network was 11.15%, and the average absolute percentage errors of equal weight method, dominance matrix method and series method were 10.16%, 10.10% and 9.69% respectively. Only when the mean absolute percentage error of the prediction results obtained by the series method is less than 10%, it can be considered that the prediction accuracy of the series method is higher and superior to other models.

If we only want to predict the national epidemic morbidity in the short term, the single ARIMa model has good prediction effect and high accuracy. However, if we want to apply it to the long-term prediction, the advantages of the series combination model are more obvious, and its overall prediction effect is the best.

VI. CONCLUSION

The disease prediction model studied in this paper has both linear and nonlinear time series characteristics. However, BP neural network just has an excellent performance in nonlinear prediction, so the combination model can make up for the shortcomings of a single model in practical application. It can also make full use of the advantages of ARIMA and BP neural networks, and greatly improve the accuracy of prediction. The experimental results show that the average absolute percentage error of the prediction results obtained by the series method in this paper is less than 10%, and its prediction accuracy is higher and better than other models. It can provide scientific references for disease prevention and control measures.

In this paper, the sample size of the training set is not large enough, and the high dimensionality of the data brings a lot of redundant information, which easily leads to over-fitting of the model and the reduction of prediction performance. To minimize the problems caused by small sample size and high dimensionality data, future work will continue to accumulate and supplement the complete samples of clinical follow-up data, expand the sample size of the training set, and use the new samples to evaluate the two classifiers constructed in this paper.

ACKNOWLEDGMENT

The study was supported by “Key project of Jining Health Commission "Public Health Monitoring and Early Warning System of Jining City" (No. SZBM-2021-D0014)”.

REFERENCES

[1] XU S, CHAN H K, ZHANG T. Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach[J]. Transportation Research Part E, 2019, 122.

[2] Witteveen E, Wieske L, Sommers J, et al. Early Prediction of Intensive Care Unit–Acquired Weakness:A Multicenter External Validation Study[J]. Journal of Intensive Care Medicine, 2020,35(6):595-605.

[3] M M K G, G A D, B R J, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration[J]. Annals of internal medicine, 2015, 162(10): 735-736.

[4] Qingui C, Lishan Z, Shanhui G, et al. Prognosis predictive value of the Oxford Acute Severity of Illness Score for sepsis: a retrospective cohort study[J]. PeerJ, 2019,7: e7083.

[5] Ho K M, Williams T A, Harahsheh Y, et al. Using patient admission characteristics alone to predict mortality of critically ill patients: a comparison of three prognostic scores[J]. Journal of critical care, 2015, 31(1):21-25.

[6] Sedloň P, Kameník L, Škvařil J, et al. Comparison of the accuracy and correctness of mortality estimates for intensive care unit patients in internal clinics of the Czech Republic using APACHE II,APACHE IV, SAPS 3 and MPMoIII models[J]. Medicinski Glasnik Official Publication of the Medical Association of Zenica Doboj Canton Bosnia & Herzegovina, 2016, 13(2):82.

[7] Ko M, Shim M, Lee S M, et al. Performance of APACHE IV in Medical Intensive Care Unit Patients:Comparisons with APACHE II, SAPS 3, and MPM0 III[J]. Acute and Critical Care (v.32;2017), 2018, 33(4): 216-221.

[8] Miechowicz J. Prognostic scoring systems for mortality in intensive care units -- the APACHEmodel[J]. Anaesthesiology intensive therapy, 2015,47(1): 46-49.

[9] Salluh J I, Soares M. ICU severity of illness scores: APACHE, SAPS and MPM[J]. Curr Opin CritCare, 2014, 20(5):557-565.

[10] Ferrando-Vivas P, Jones A, Rowan K M, et al. Development and validation of the new ICNARCmodel for prediction of acute hospital mortality in adult critical care[J]. Journal of Critical Care, 2016, 38:335-339.

[11] Hadique S, Culp S, Sangani R G, et al. Derivation and Validation of a Prognostic Model to PredictSix-Month Mortality in an Intensive Care Unit Population[J]. Annals of the American thoracic society, 2017, 14(10): 1556-1561.

[12] L Z, W X, Y W. Prognostic scoring systems for mortality in intensive care units — the APACHEmodel[J]. Research Square, 2020.[183] A. S. Hauser, M. M. Attwood, M. Rask-Andersen, et al. Trends in GPCR drug discovery: new agents, targets and indications[J]. Nature Reviews Drug Discovery, 2017, 16(12): 829-842.

[13] Y. P. Chen, Y. Q. Wang, J. W. Lv, et al. Identification and validation of novel microenvironment-based immune molecular subgroups of head and neck squamous cell carcinoma: implications for immunotherapy[J]. Ann Oncol, 2019, 1(30): 68-75.

[14] B. Li, Y. Cui, D. K. Nambiar, et al. The Immune Subtypes and Landscape of Squamous Cell Carcinoma[J]. Clin Cancer Res, 2019, 12(25): 3528.

[15] D. Bruni, H. K. Angell, J. Galon. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy[J]. Nature Reviews Cancer, 2020, 20(11): 662-680.

[16] Y. Zhang, M. Yu, Y. Jing, et al. Baseline immunity and impact of chemotherapy on immune microenvironment in cervical cancer[J]. Brit J Cancer, 2021, 124(2): 414-424.

[17] R. S. Herbst, J. C. Soria, M. Kowanetz, et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients[J]. Nature, 2014, 515(7528): 563-7.

[18] R. R. Ji, S. D. Chasalow, L. Wang, et al. An immune-active tumor microenvironment favors clinical response to ipilimumab[J]. Cancer Immunology, Immunotherapy, 2012, 61(7): 1019-31.

[19] S. Wang, Z. He, X. Wang, et al. Antigen presentation and tumor immunogenicity in cancer immunotherapy response prediction[J]. eLife, 2019, (8): e49020.

[20] M. O'Hayre, M. S. Degese, J. S. Gutkind. Novel insights into G protein and G protein-coupled receptor signaling in cancer[J]. Current Opinion in Cell Biology, 2014, (27): 126-35.

[21] D. Balli, A. J. Rech, B. Z. Stanger, et al. Immune Cytolytic Activity Stratifies Molecular Subsets of Human Pancreatic Cancer[J]. Clin Cancer Res, 2017, 12(23): 3129.

[22] J. Galon, D. Bruni. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies[J]. Nature Reviews Drug Discovery, 2019, 3(18): 197-218.

[23] D. Sia, Y. Jiao, I. Martinez-Quetglas, et al. Identification of an Immune-specific Class of Hepatocellular Carcinoma, Based on Molecular Features[J]. Gastroenterology, 2017, 153(3): 812-826.

[24] A. Mayakonda, D. C. Lin, Y. Assenov, et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. R Package Version 2.4.05[J]. Genome research, 2018, 11(28): 1747-1756.

[25] F. Martínez-Jiménez, F. Muiños, I. Sentís, et al. A compendium of mutational cancer driver genes[J]. Nature Reviews Cancer, 2020, 20(10): 555-572.

Evaluation of Spiral Pattern Watermarking Scheme for Common Attacks to Social Media Images

Tiew Boon Li[1], Jasni Mohamad Zain[2], Dr. Syifak Izhar Hisham[3], Alya Afikah Usop[4]
Faculty of Computing, Universiti Malaysia Pahang, Lebuhraya Tun Razak, Pahang[1, 3]
Institute for Big Data Analytics and Artificial Intelligence, Kompleks Al-Khawarizmi, UiTM, Shah Alam, Malaysia[2]
Faculty of Computing, Universiti Malaysia Pahang, Pekan, Pahang[4]

***Abstract*—The 21st century might be considered the "boom" period for social networking due to the fast expansion of social media use. In terms of user privacy and security regulations, a plethora of new requirements, issues, and concerns have arisen due to the proliferation of social media. With the increase in social media use, images on social media are often modified or fabricated for certain purposes. Therefore, this work implements and evaluates the SPIRAL-LSB algorithm for common attacks for social media images. Image compression was also discussed as images published to social media platforms was often compressed. An analysis was performed to assess the algorithm's output on social media images. The experiments were carried out prior to and after uploading to the Instagram platform. The dataset was subjected to image splicing, copy-move, cut-and-paste, text insertion, and 3D-sticker insertion attacks. The outcome of SPIRAL-LSB was effective for text insertion attacks solely. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) were selected as the experiment's metrics. The average PSNR value is 63.25, and the SSIM value is 0.99964, both of which are regarded high. This indicates that the watermark has not degraded the quality of the images. This work was designed for usage on social media for intellectual property reasons and may be used to validate the validity of social media images and prevent issues with image integrity, such as image manipulation.**

Keywords*—Spiral pattern; fragile watermarking; social media; LSB substitution

I. Introduction

The application of social media is rapidly intensifying, and the twenty-first century may be defined as the "boom" time for social networking. According to Smart Insights data, there were approximately 3.484 billion social media users in February 2019. The Smart Insight survey reported the number of social media users is increasing by 9% every year, and this trend is expected to continue [1]. Currently, the social media users symbolize 45% of the worldwide population [2]. The most frequent users of social media are digital natives, a group of people who were born or grew up in the digital age and are familiar with numerous technologies and systems, and the Millennial Generation, individuals who became adults around the turn of the twenty- first century.

Moreover, according to [3], the usage and sharing of information through the internet is an inherent or intrinsic element of university students' lives. The study's results indicate that students often use Facebook to share information. Numerous individuals disclose their personal information without thinking of the consequences. Consequently, social media platforms have developed into a vast repository of sensitive data. Users are more receptive to friend invitations and trust goods sent to them by friends [4].

The move toward visual social media is being pushed in part by changes in social media user behaviors as a result of the enhanced mobile internet experience. Due to the widespread use of advanced software applications such as Picasa and Photoshop, image manipulations have become a fairly popular and effortless action for everyone. Edited images are often aesthetically appealing and difficult to differentiate from unaltered ones. There is a growing tendency toward the use of modified images in every aspect of our daily lives, such as news reporting, blogging, and advertising [5]. This often leads to user deception [6] which has the potential to influence and manipulate public opinion, ranging from teens' self-esteem and personal health choices to public opinion in significant political areas.

Although manipulated images are often uncovered, it may take weeks, and by that time, millions of people's opinions have already been influenced. This may raise severe concerns about the trustworthiness of digital multimedia, since it puts questions on the face value of the information we receive on a regular basis through the Internet [7]. This issue is getting more severe, presenting major difficulties to society. Revolution of Internet and technology enables pirates to unlawfully utilize the features to manipulate images [8]. Thus, the necessity for digital media authentication techniques becomes vital to ensuring that work is not tampered with, particularly in crucial circumstances such as social media politics, medical safety, internet banking, military data transmission, and forensic investigations.

In disciplines such as forensics, medical imaging, and military and industrial images, the integrity of a digital image is critical [9]. Digital watermarking is considered a technological category in dealing with integrity issues [10]. Hence, to preserve social media images and identify ownership, digital watermarking is essential. Without watermarks, images on social media are vulnerable to theft and illegal use [11]. In theory, digital watermarking can distinguish between various sorts of third- party manipulations and attacks.

A. Integrity and Authentication of Digital Images

In between the techniques for securing digital data, digital watermarking has grown in popularity among academicians and users due to its variety and ability to retain the integrity

and authenticity of digital images. The term "image authentication" refers to the process of determining the legitimacy of digital images. Among the methods for establishing image authenticity are location of tampering. As mentioned in the previous section, digital watermarking may potentially discriminate between different types of manipulations and assaults by a third party. Manipulations in this instance include those that are permitted and those that are not permitted [12].

There are three techniques to watermarking that includes fragile watermarking, robust watermarking, and semi-fragile watermarking, which combine fragile and robust aspects. Watermarking images is critical for preserving personal data privacy and avoiding image tampering [10]. In general, an image authentication technique is composed of two stages: embedding and validation. The embedding stage embeds the authentication data in an image and stores it as proof of the image's validity; the validation stage compares two images: one evaluated for the watermarked image, and another extracted from the watermarked image and determines whether the image has been modified or not [13].

Authentication through fragile watermarking is performed by embedding a watermark into the image, which is quickly altered or destroyed when the watermarked image is manipulated or attacked. When compared to the image's real content, the presence or absence of the watermark is identified [12]. Several prominent strategies allow for the localization and recovery of changed regions in a block-wise manner. While embedding, certain techniques may provide metadata about the image. In contrast, systems based on robust watermarking assume that a good watermark is impervious to image manipulations.

Digital watermarking, among current approaches and owing to its exceptional qualities, is an efficient option for protecting multimedia data in a variety of industries. The primary benefit of digital watermarking is that the authentication data is included directly in the image data. The authentication information is preserved, even if the watermarked image is converted to a different format and the retrieval procedure is described as simpler and less complex [14]. Table I shows the key contrasts between these three notions namely cryptography, steganography, and watermarking. These three methods are commonly used as data security techniques. Fig. 1 shows the data security techniques.

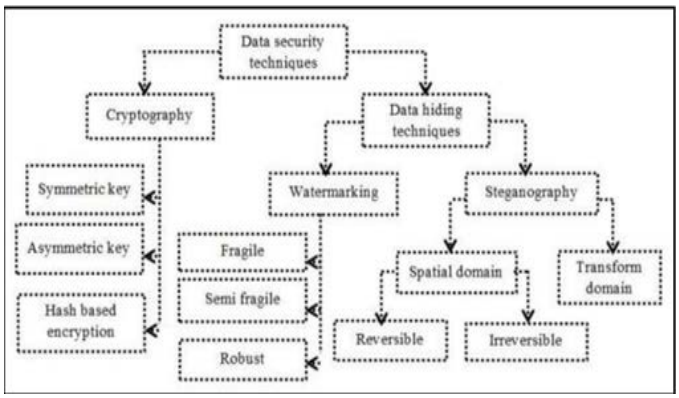


Fig. 1. Data Security Techniques [14].

TABLE I. COMPARISON OF CRYPTOGRAPHY, WATERMARKING AND STEGANOGRAPHY [14]

Criterion	Cryptography	Watermarking	Steganography
Objective	Encrypted communication	Content authentication and copyright preservation	Covert communication
Authentication	Yes	Yes	No
Cover selection	Not required	Usually image, audio, or video	Any digital object
Key	Mandatory	Optional	Optional
Attacks	Cryptanalysis attacks; Ciphertext only attacks; Known-plaintext attack; Chosen-plaintext attack; Brute-force attack; Man-in the-middle attack; Birthday attack; Timing attack; Dictionary attack.	Image processing attacks; Salt and pepper noise; Cropping attack; Rotation attack; Sharpening attack; JPEG attack; Median filtering attack; Quantization; Temporal modification.	Steganalysis attacks; Regular and singular analysis; Pixel difference histogram attack; Chi-square attack; Sample pair analysis.
Robustness	Not required	Should be high	Should be high
HC	Not required	Should be high	Should be high
Imperceptibility	Not required	Should be high	Should be high
Visibility	Always visible	Depending upon the type of watermarking, it can be visible or invisible	Always visible
Output	Encrypted text	Watermarked object	Camouflage object
Merits	It offers both authentication and integrity, along with confidentiality	It offers both authentication and integrity, along with confidentiality	None apart from the sender and receiver can suspect the existence of the communication
Demerits	The communication is visible to the outsider	HC is usually low	Steganography itself alone cannot provide authentication and integrity
Purpose is lost	If the communicating message is decrypted	If the watermark is abolished or heavily tampered	If the attacker knows communication
Origin	Very ancient	Modern era	Very ancient

The remainder of this paper is organized as follows: Section 2: Literature Review, Section 3: Methodology, Section 4: Results and Discussion, Section 5: Conclusion, Section 6: Acknowledgement and Section 7: References.

II. LITERATURE REVIEW

A. Related Works

An overview of fragile watermarking systems for image authentication is presented by [15]. The limited embedding capability and amount of tampering are two major challenges

that motivate study in this field. This review covers the overall framework of the fragile watermarking system, as well as the many types of assaults and parameters used to evaluate the methods. The researchers will be able to quickly analyze current achievements in this field by using comparative analysis and quantitative comparisons of fundamental schemes and their variants with enhancements.

The authors [16] propose a secure fragile image watermarking system that is used to identify image content alteration or manipulation. The proposed approach consists of two steps: computing a secure authentication code/watermark bit from some of each pixel's most significant bits, and then hiding the watermark bit in the least significant bit (LSB) of each pixel using a recommended watermark embedding procedure. On a series of grayscale images, the proposed watermarking method is evaluated, and the watermarked image's quality is shown.

The authors [17] present a dual watermarking technique capable of integrating authentication, copyright protection, and image recovery functionalities into a single cover image. The robust scheme protects against copyright infringement by utilizing a single watermark in the discrete cosine transform (DCT) domain, whereas the fragile scheme protects against copyright infringement by utilizing two self- embedding watermarks in a spatial domain for authenticating and restoring digital image content.

The authors [18] presented a new technique for copyright protection, data security and content authentication of multimedia images. The authentication of the content has been ensured by embedding a fragile watermark in the spatial domain while copyright protection has been taken care of utilizing a robust watermark. The fragile watermark embedding makes the system capable of tamper detection and localization with average value more than 45% for all signal processing and geometric attacks. The average Peak-Signal- to-Noise Ratio (PSNR) achieved for both schemes are greater than 41 dB.

The author [19] developed a unique spiral numbering pattern for fragile digital watermarking schemes. The developed scheme is designed to achieve a good numbering pattern, exact detection, and image recovery. The limitations of the proposed scheme are works on gray-scale images only and square images.

To address the identified gap in the watermarking literature, the majority of studies have been conducted on medical images; however, there are a few studies that have been conducted on the security of social media images via digital watermarking, and the aforementioned studies have their own limitations and weaknesses. Thus, this work implemented and evaluated a fragile watermarking method on social media images. Initially, this algorithm was shown to function for medical images but has not been demonstrated to work for social media images. Thus, our effort adds to the security and integrity of images shared on social media platforms such as Instagram.

B. Popular Social Media Platforms

- Facebook: Facebook is a large social networking website where users may share comments, photos, and links to news or other relevant items on the web, as well as live chat and watch reels. Shared information may be made publicly available or restricted to a small group of friends or family members, or to a single individual. Since its inception on February 4, 2004, Facebook has grown to over 1.59 billion monthly active users, making it one of the finest platforms for connecting people from all over the globe.

- Twitter: Twitter is ranked as one of the top social networks in the world by active users. Twitter has 192 million marketable daily active users and gains 5 million daily users in the fourth quarter of 2020 [20]. Twitter gains 5 million daily users in Q4, Projects 20 Twitter is a popular social media site because it is personal and rapid. Twitter combines instant messaging, blogging, and texting, but with brevity and mass appeal. Most people nowadays have Twitter accounts including celebrities who use Twitter to engage with followers.

- Instagram: Instagram is one of the most popular social media platforms in the modern day. Without Instagram, it is difficult to run an effective social media marketing strategy. As an image and video-centric social network, Instagram gained popularity due to its easy filter tool, which can instantly transform any shot into a high-quality one. Live video, Instagram TV-IGTV, geotagged posts, hashtags, stories, and advertisements all appear as attractive features for users. Hence, the site has around 400 million active users and was acquired by Meta in 2012. Most people utilized Instagram to share information on travel, fashion, nutrition, and craftsmanship.

- WhatsApp: WhatsApp is a cross-platform instant messaging application available on smartphones, tablets, and personal computers. This program requires an Internet connection in order to transmit photos, text, documents, audio, and video messages to other users who have installed the app on users' devices. WhatsApp Inc. was founded in January 2010 and was acquired by Meta on February 19, 2004, for about $19.3 billion. Today, over a billion people use the internet to communicate with their friends, families, and even customers.

- Snapchat: Despite the competition from other social media platforms, Snapchat continues to be one of the most popular social media platforms today, especially among younger users. Indeed, in 2021, Snapchat had approximately 428 million users worldwide. Snapchat initially used it for private image sharing, video, and messaging, as well as generating caricatures like Bitmoji characters and sharing a chronological story with users' followers.

- Reddit: Reddit is a community-driven news website where users may produce and share content. The reason users of Reddit are attracted to the site is the promise of high-quality material. Reddit members are very active and often publish something fresh and intriguing. Reddit was one of the most popular mobile social applications in the United States as of June 2021, with around 48 million monthly active users.

C. Image Compression on Social Media

Working with larger photos with a higher bit depth, the images become too enormous to send over a regular network connection. To show an image in a fair length of time and utilize a reasonable amount of space to retain the image, approaches to minimize the image's file size must be used. These approaches analyze and compress visual data using mathematical algorithms, resulting in reduced file sizes. This is known as compression.

There are two types of image compression methods: lossy and lossless. Both systems conserve storage space, but the strategies used are different. Lossless compression expresses data in mathematical formulae while retaining all the original image's information. The integrity of the original image is preserved, and the decompressed image output is bit-for-bit identical to the original image input [21].

Lossy compression shrinks files by removing extra image data from the original image. It eliminates features that are too fine for the human eye to distinguish, resulting in close approximations of the original image, but not a perfect reproduction [21]. One of the most apparent advantages of lossy compression is that it results in a much lower file size compared to lossless compression, but at the expense of quality. With lossy compression, it is necessary to establish a compromise between file size and image quality. As shown in Fig. 2, with 50 percent compression, we reduced the size of the image file by 90 percent. With a compression ratio of 80%, we were able to reduce the image file size by 95%.

Lossless compression, on the other hand, is the process of reducing the size of an image without compromising its quality. Typically, JPEG and PNG files are stripped of unnecessary information. Lossless image formats include RAW, BMP, GIF, and PNG. With small reductions in image file sizes, there is no loss of image quality. Fig. 3 depicted the original and lossless compressed image.

Huge volumes of fresh data are constantly posted to Instagram's servers as a result of the millions of new posts submitted daily. The problem might soon spiral out of control if terabytes of data are uploaded every day. Instagram compresses both image and video postings to decrease server strain and maintain a steady flow of content. The user experience is also a factor in the compression. Some large videos and images would take a long time to upload if compression were not available. Users may be dissuaded from uploading further data if there are lengthy wait periods. In turn, this would result in decreased Instagram traffic and user engagement. Instagram has effectively avoided this problem, whether on purpose or not, by imposing rigorous limits and limitations on image sizes.

Fig. 2. Degree of Lossy Compression [22].

Fig. 3. Original Image and Lossless Compressed Image [22].

The issue arises due to Instagram's excessive JPEG compression of uploaded and shared images. JPEG employs lossy compression, which discards data, increasing the likelihood that watermarked data may be discarded. When users upload JPEGs to Instagram, they are compressed again, but by Instagram. In essence, users are double the compression and sacrificing quality. PNG uses lossless compression and hence should be less impacted by Instagram's. Uploading images in PNG format is advised to maintain a small size and good quality of images.

D. Common Attacks on Social Media Images

Digital images may be manipulated or attacked to deceive by changing some of the image's critical information. These attacks can be performed on social media images and lead to negative consequences such as financial loss, business fraud, defamation and to serious extent, cybercrime proceedings. These alterations are extremely destructive to some critical images, such as military and medical images, and such images should be preserved. The authors [23] categorized image forgery techniques into two basic approaches: active and passive, as seen in Fig. 4.

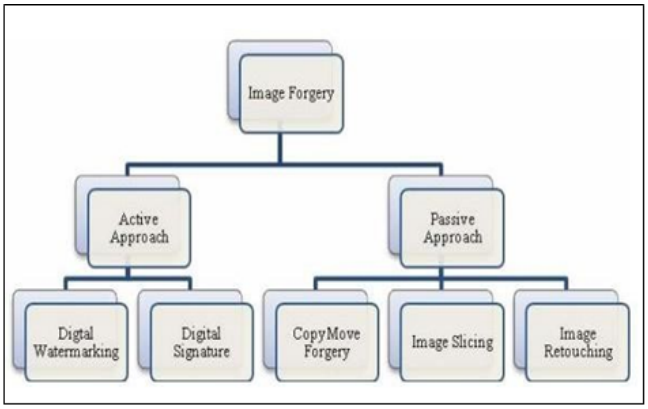


Fig. 4. Image Forgery Techniques.

1) Image splicing: Image splicing to generate counterfeit photos is more aggressive than image editing. Image splicing is a fundamentally basic procedure that may be done as crops and pastes regions from the same or other sources. This approach refers to a paste-up done by gluing together images utilizing digital tools available such as Photoshop.

In image splicing method there is composition of two or more photos, which are merged to generate a false image. Examples include some notorious news reporting situations involving the use of falsified images. Fig. 5 illustrates how to generate a forge image; by transferring a spliced piece from the source image into a target image, it creates a composite image of scenery which is a forge image.

The authors [24] describe image splicing as a collage created by adhering photographic images together. Image splicing is a method that combines two or more images to generate a new fictitious image. The image splicing technique is more aggressive than the resampling technique [24]. It is often followed by post processing such as blurring, compression, and scaling. It is often employed as the first stage in photomontage, a technique that is very popular in digital image content modification.

The authors [25] identified an image splicing approach that is based on image texture analysis, which defines image portions based on their texture richness. The texture content of an image is used to describe it in this manner. The modified image created by splicing might be utilized in news stories, photography contests, or as major evidence in academic papers, which could have a decisive impact.

2) Copy move attack: The copy move forgery is one of the commonly utilized forms of image manipulation method. In this approach, one has to cover a section of the image in order to add or delete information. In a copy-move attack, the objective is to disguise anything in the original image with some other section of the same image. The example of copy-move type is as shown in Fig. 6 when a troop of soldiers are cloned to cover George Bush.

The authors [24] claimed that copy move attack is when a portion of an image is copied and pasted into different locations within the same image to conceal information or change the meaning of the image. The digital image copy-move forgery technique involves the repetition ozone or more areas at various positions inside the same image. Frequently, duplicated portions are extended, shrunk, or rotated to increase the convincingness of forgeries, making it more difficult to identify forgeries.

3) Image retouching: Previously, retouched images were intended for magazine covers and mostly used on celebrities. The advancement of technology has increased the ease with which images may be retouched, resulting in a rise in over-perfect images. For example, Zendaya has taken to Instagram to criticize publications for retouching magazine figures as seen on Fig. 7.

Most alarming consequence of image retouching is the booming of selfie culture, which promotes a society preoccupied with money, beauty, power, and fame. Photoshop and Beauty Camera paving the way for unattainable beauty standards and are thereby contributing to the rising pandemic of body dysmorphia and mental health problems among today's youth. The image is not drastically altered during image retouching, but some characteristics of the image are enhanced or diminished, a technique that is quite common in the majority of photo editing software. In most image magazines, there is a need for image attractiveness, which results in the enhancement of some aspects of an image, oblivious to the fact that such approach is illegal.

4) Meme manipulation: The term "meme" derives from the Greek "mimesis," which refers to the way art imitates life [26]. Memes have been used as a weapon in cultural battles for more than a decade. Memes are more convincing than most people believe. On a social media timeline, a well-placed meme might lead down a rabbit hole of radicalization, misinformation, and extremism. In this scenario, Internet Memes stepped in as a compelling tool for users to express themselves in the ironic format, which often combines visual and text materials. Fig. 8 shows an example of a political meme between North Korea and America.

The authors [27] define Internet memes as artifacts of participatory digital culture, an excellent description of the functional purpose. Memes have the capacity to be made, utilized, spread, and remixed by anybody with Internet connection creating previously unimaginable opportunities for engagement in social and political concerns. To date, research on memes has been concerned with their contribution to the expression of political ideas and of subcultural identity [28].

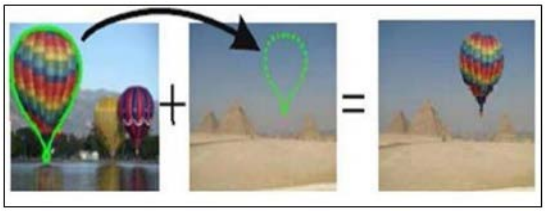


Fig. 5. Image Splicing.

Fig. 6. Copy Move Forgery Image.

Fig. 7. Zendaya Retouched Image.

Fig. 8. Political Meme.

III. Methodology

In the image authentication phase, collected sample images are input into the algorithms and output are obtained. Data collection starts the flow of the work. 15 original colored-images are acquired which acts as the host image. Host image was fed in the algorithm to be embedded with a watermark. The authentication watermark was a 2-bit authentication watermark, intended to compare the intensity (v) and the parity bit (p) for the detection of tamper in the colored-image. Following that, the host image undergoes block division to produce an image block (in pixels) using block numbering in a spiral pattern. After the embedding process, a watermarked image is produced.

For the purpose of testing, the watermarked image was manipulated with five different attacks, namely, image splicing, copy-move forgery, cut-and-paste, sticker insertion and text insertion. These five attacks are the most common attacks performed on social media images. The 15 sample images were manipulated with each type of attack, thus producing 75 attacked images as the input to the algorithm. To depict the image compression influence on social media images, the image authentication process was performed twice, first prior to the upload into social media and second after uploaded into social media.

The functional block diagram for watermark numbering, mapping, generation, and embedding was shown in Fig. 9. The technique in numbering is in a spiral manner. The following algorithms describe how the 2-tuple watermark of each sub-block was generated and embedded, which adapted from [19]:

1) Set the LSB of each pixel within the block of B to zero.

2) Calculate the average intensity of the block, AvgB and each of its sub-blocks, AvgBs, respectively.

3) Generate the authentication watermark, v, of each sub-block. V is 1 if the AvgBs is bigger than AvgB or 0 if otherwise,

4) Generate the parity check bit, p of each sub-block. P is 1 if the parity number is odd, and 0 if otherwise.

5) Obtain the original image, A, from the mapping sequence done at the first phase.

6) Compute the average intensity of each sub-block again within A, AvgAs.

7) Embed the 2-tuple watermark (v, p) each in one LSB of each pixel in Bs.

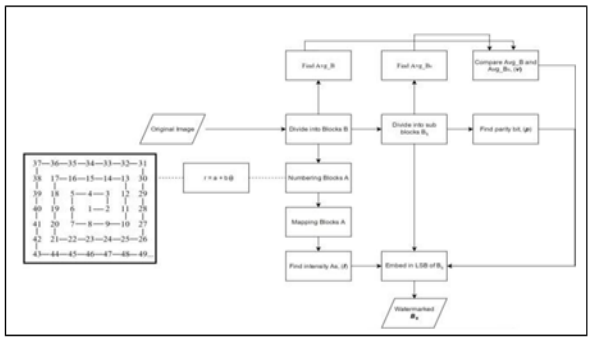


Fig. 9. Embedding Process [19].

In the SPIRAL-LSB scheme, two levels of detection phase were developed to guarantee no missing tamper when detecting. The first level would examine the parity bits and values of the average intensity in the sub-blocks, while the second level would examine the parity bits and values of the average intensity in the blocks containing the sub-blocks examined on the first level. This is done to ensure a high detection rate.

The experimental images were initially separated into non-overlapping 8 by 8-pixel blocks, similar to the watermarking embedding procedure. For each Br block, the LSBs of each Br pixel were set to zero and its average intensity, designated by Avg_Br, was computed. Then, a two- level detection was conducted. The procedure of hierarchical tamper detection scheme from [29, 30] is outlined below:

- Level 1 detection: For each 4 × 4-pixel sub-block Brs inside the block Br, do the following operations:

1) Extract v and p from Brs.

2) Set the LSBs of each pixel within each Brs to zero and compute the average intensity for each sub-block Brs, denoted as avg_Brs.

3) Set the algebraic relation v'=1 if avg_Brs >= avg_Br, otherwise, set it to 0.

4) Calculate the total number of 1s in avg_Brs and denote it as Ps.

5) Set the parity check bit p' of Brs to 1 if Ps is even, otherwise, set it to 0.

6) Compare p' with p and compare v' with v. If unequal, mark Brs as tampered and complete the detection for Brs; otherwise mark it as valid.

- Level 2 detection: For each valid 8x8 pixel block Br, do the following operations:

1) Search the block number of block C, where block C is the one in which the intensity feature of block Br is embedded.

2) Locate block C.

3) If block C is marked tampered, assume block Br is valid and complete the test.

4) If block C is valid, perform the following steps:

a) Get the 7-bit intensity of each Brs by extracting the LSBs from each pixel in the corresponding block within block C, padding one zero to the end to make an 8-bit value.

b) Compare with avg_Brs and mark Br tampered if they are different.

IV. Results and Discussion

After The samples to test the SPIRAL-LSB scheme were in PNG and JPG format with RGB colored type. The images were in square-sizes. From the algorithm applied, our result showed that embedding scheme with the block spiraling and starting the numbering in the middle would produce significant PSNR values, which as a whole, we can say all were above 55 dB, with average of 65.09 dB reported from the output data of 15 samples.

The highest value was 67.5 dB and lowest was 58.98 dB. Fig. 10 depicted the graph of the recorded PSNR values. Moreover, the SSIM value produced a correlation average value of 0.99964 which we regarded as very high. The produced SSIM value corresponds to one, indicating that the watermarked image closely resembles the original. The highest and lowest values were 0.9992 and 0.9998, respectively. Fig. 11 shows the graph of the recorded SSIM value.

A. Text Insertion Attack

In Fig. 12, a text "Vaccinated!" was inserted on the image (a) to produce tampered image (b). After acquiring the tampered image, it was tested prior to uploading it to Instagram. Figure (c) is the result before uploading while figure (d) shows the result after uploading. The tampered region is detected in red color. The tamper was detected and marked it in red, as shown in Fig. 12(d).

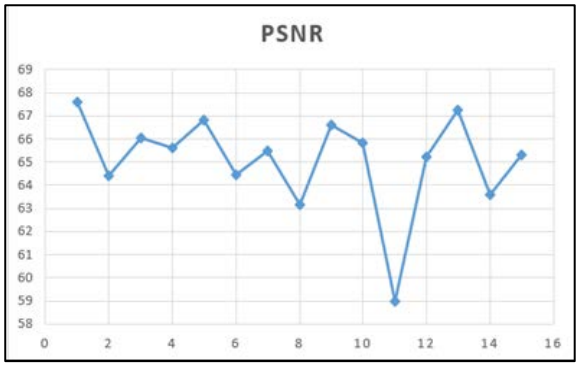


Fig. 10. PSNR Value.

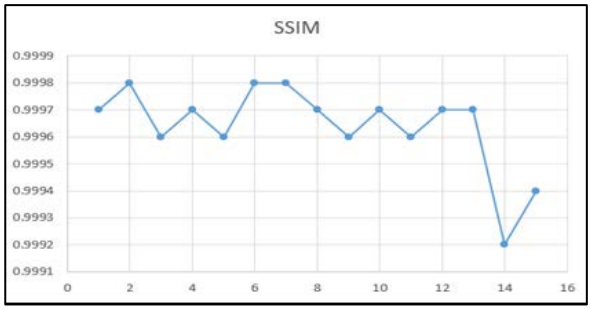


Fig. 11. SSIM Value.

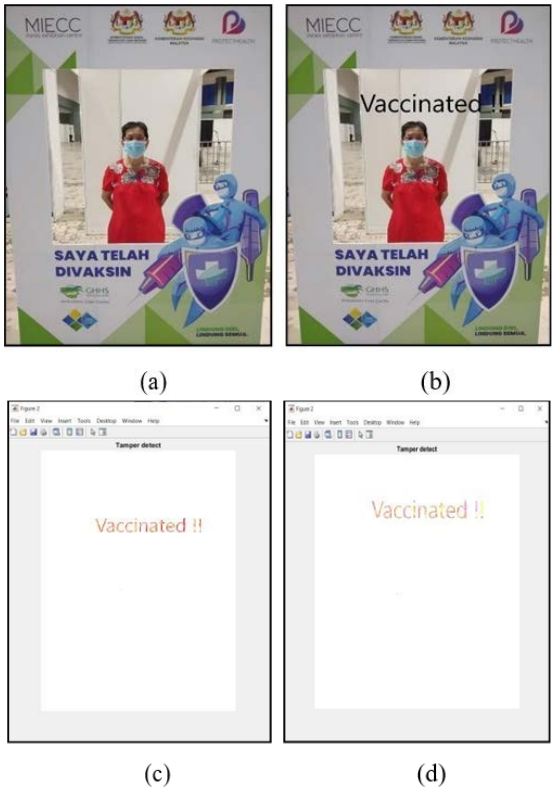


Fig. 12. (a) Original Image, (b) Tampered Image (c) Before Post, (d) After Posted.

In Fig. 13, a text "Sail boat" was inserted on the image (a) to produce a tampered image (b). After acquiring a tampered image, it was tested prior to upload in Instagram. Figure (c) is the result before uploading while figure (d) shows the result after uploading. Figure (a) was a general image taken from an image database, so the result shows no noise detected even after being uploaded to social media. The tamper was detected and marked it in red, as shown in Fig. 13(d).

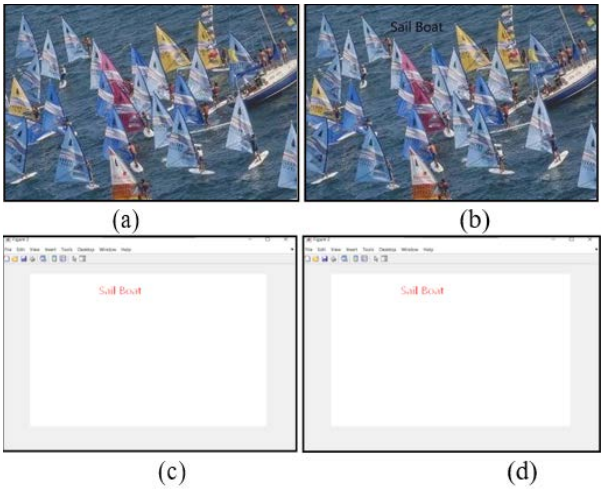
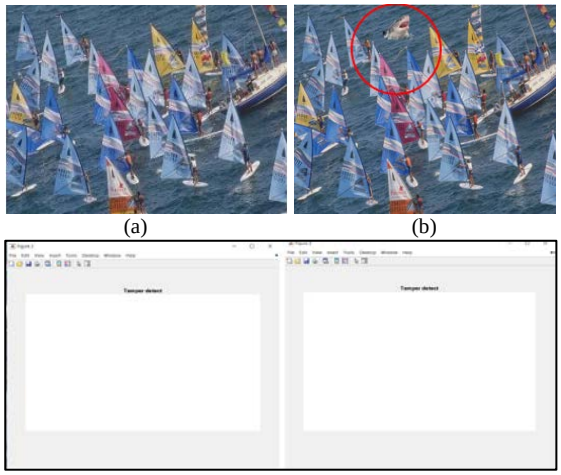


Fig. 13. (a) Original Image, (b) Tampered Image, (c) Before Post, (d) After Posted.

B. Image Splicing Attack

For image splicing attacks, Fig. 14(b) shows a spliced image. The objects were circled in red. Fig. 14(c) displays the results before uploaded into Instagram while Fig. 14(d) depicts the results after uploaded into Instagram. The results show that the tampered regions failed to be detected. JPEG compression applied by Instagram for uploaded images has not greatly affected the performance of the algorithm, since all the experiments provide identical result prior upload and after uploaded to Instagram platform.



(a) (b)

(c) (d)

Fig. 14. (a) Original Image, (b) Spliced Image, (c) Before Post, (d) After Posted.

Various attacks were done based on the studies revealing the common attacks to social media images. From the findings of the output, we can deduce that SPIRAL-LSB is able to detect text insertion attacks exclusively. Although some detection results display little fading after uploaded into social media. However, watermark is still intact with the host image as detection of tamper is successful. From the text insertion output, the deduction of the scheme was robust against JPEG compression because the tampered region is detected clearly after posted to Instagram.

From the 15 images that have image splicing attack, all of them were unable to be detected regardless of compression issue. Copy-move, cut-and-paste, and 3D-sticker insertion attacks produced identical outcomes. Hence, we can conclude that SPIRAL-LSB is inefficient to detect tamper for image splicing, copy-move, cut-and-paste and 3D-sticker insertion attacks. The watermarking scheme failed to detect the tampered region before the images posted to Instagram. Thus, SPIRAL-LSB is suitable for social media uploaded images that have been attacked by text insertion. For image splicing, copy-move, cut-and-paste and 3D-sticker insertion attacks, it is not suitable to be used.

The failure to detect for copy-move attack might be SPIRAL-LSB scheme was using comparison of intensity and parity bits to detect whether there was any tamper in the image or not. However, image splicing, copy-move, cut-and-paste and 3D-sticker insertion attacks may need the use of another approach to identify and key point-based forgery detection method have been proven helpful in detecting copy-move forgeries [31]. For image splicing, using two Markov features: coefficient-wise Markov features and block-wise Markov features in the discrete cosine transform (DCT) domain produce high detection accuracy [32]. Thus, SPIRAL-LSB was not effective to detect image splicing attack.

In this work, we used Least Significant Bit (LSB) which is a spatial domain technique. According to [33], the embedding of the watermark into the original image is done by selecting a subset of pixels and substituting the least significant bit of the selected pixels with the watermark bits. The LSB techniques, are easy to implement and requires a little computation cost for both embedding and extraction processes. On the hand, they are sensitive to signal processing operations and generally show reduced robustness to different attacks. Even though there are a large number of suggested LSB algorithms, there is still a lack of a robust solution, necessitating further study in this field.

Spatial domain techniques are simple and have a high payload, work directly on the pixel level, but these are not robust against various attacks [34]. In spatial domain the information is added simply by just varying the pixel values of the host signal. The values of some colors or pixels are also directly editable in the spatial domain techniques. In the least significant bit (LSB) substitution technique, the watermark is added in the least significant bit of each pixel. When the extraction of information is needed, the LSB of each pixel is read. However, the major disadvantage of this watermarking is that it is not robust again various attacks according to [35]. So, the weakness of least significant bit technique is shown in the experiments in this study. SPIRAL-LSB could not detect the tampered regions of image splicing, copy-move, cut-and-paste and 3D-sticker insertion attacks.

In our experiment, the performed attacks can be considered as pixel level tampering. Thus, SPIRAL-LSB algorithm is a block-wise technique, and it cannot detect pixel-level tampering. This drawback is called a localization problem and it was reported by [36] in 2002. Subsequently, fragile watermarking techniques have been developed to address localization problem [37, 38]. Recently, the authors [39] proposed two related fragile watermarking techniques. The first method is a statistical technique which is capable of detecting pixel-level tampering if the tampered area is small. The second one improves the tamper detection capability for a larger area by incorporating a hybrid of block-wise and pixelwise mechanism. However, the use of block information reduces its tamper resistance capability.

From the previous research done by other researchers, it is proven that LSB substitution techniques have weaknesses and limited robustness under various attacks such as lossy compression which implemented by Instagram sites. Instagram uses a lossy compression technique (JPEG compression) that reduces the image's quality and size to save storage space, reduce the amount of computing resources required for image processing, and speed up the loading or display of an image on a user's timeline. In comparison to the original image, the image posted on social media contains distortion and noise. Thus, the watermarking scheme also detected noise which is in yellow color in posted images.

As it worked in a spiral manner, which started at the center, the image processed should be in square size to ensure all the blocks were numbered. The scheme could only number the image blocks in the square which also led to generating the watermarking data in the square too, not in total if the image were in a rectangle shape. This limitation made the scheme not compatible with other social media sites images such as Facebook as Facebook support images vary in sizes.

V. Conclusion

Social media attacks represent the largest modern threat vector and are at all-high because roughly 3.5 billion people are on social media. Image splicing, copy-move, cut-and-paste, text, and 3D-sticker insertion were the most common types of attacks on social media. Social media platforms are often used for authentication to other website, applications, thus, this is a major attack vector. It can also be used to compromise various sectors for damage to reputation, operation, and financial gain. Hence, authentication on social media images is needed to protect the integrity of images.

This research has demonstrated that watermarking can provide authenticity for social media images. The fragile watermarking techniques for authentication with unique numbering, SPIRAL-LSB have been devised. This research has proven the existing techniques in fragile watermarking of color images by offering a way to embed in LSB in each plane of RGB without having the problem of less space or high data capacity. SPIRAL-LSB offers a novel way to number the blocks of the original image before being mapped while embedding. The spiral scan allows the data to be located farther and the operation time to be short. Although the watermarking scheme is only effective on text insertion attacks, it is proven to be robust against the effect of applying lossy compression, for instance JPEG, to such images. Despite the completion of this project, the necessity for more improvement in the future is required as the world is going through changes.

Acknowledgment

This research work is supported by a grant entitled ‘Authentication Watermarking in Digital Text Document Images using Unique Pattern Numbering and Mapping’ (RDU190366) and PGRS200369 supported by Universiti Malaysia Pahang.

References

[1] Barrett-Maitland, N., & Lynch, J. (2020). Social Media, ethics, and the privacy paradox. Security and Privacy from a Legal, Ethical, and Technical Perspective. https://doi.org/10.5772/intechopen.90906.

[2] Bullock, L., Gurd, J., & Hanlon, A. (2022, June 1). Global Social Media Statistics Research Summary 2022 [June 2022]. Smart Insights. Retrieved June 21, 2022, from https://www.smartinsights.com/social-media- marketing/social-media-strategy/new-global-social-media-research.

[3] Khan, M. N., Ashraf, M. A., Seinen, D., Khan, K. U., & Laar, R. A. (2021). Social Media for knowledge acquisition and dissemination: The impact of the COVID-19 pandemic on Collaborative Learning Driven Social Media Adoption. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.648253.

[4] Lenhart, A. (2019, December 31). Chapter 4: social media and friendships. Pew Research Center: Internet, Science & Tech. Retrieved June 21, 2022, from https://www.pewresearch.org/internet/ 2015/08/06/chapter-4- social-media-and-friendships/.

[5] Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2017, November 6). Advances in social media research: Past, Present, and Future - Information Systems Frontiers. SpringerLink. Retrieved June 21, 2022, from https://link.springer.com/article/ 10.1007/s10796-017- 9810-y.

[6] Memon, A. M., Sharma, S. G., Mohite, S. S., & Jain, S. (2018, November). The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature. Indian journal of psychiatry. Retrieved June 21, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6278213/.

[7] Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A., & Wang, Y. (2020, July 10). Setting the future of digital and social media marketing research: Perspectives and Research Propositions. International Journal of Information Management. Retrieved June 21, 2022, from https://www.sciencedirect.com/ science/article/pii/S026840122 0308082.

[8] Linas Jurkauskas. (2015, June). Digital piracy as an innovation in the recording industry - AAU. AALBORG UNIVERSITY. Retrieved February 8, 2022, from https://projekter.aau.dk/projekter/files/213765961/Digital_Pira cy_as_an_Innovation_in_Recording_ Industry_by_L._Jurkaus kas.pdf.

[9] Begum, M., & Uddin, M. S. (2020). Digital Image Watermarking Techniques: A Review.

[10] Cox, I. J., Miller, M. L. and Bloom, J. A. 2002. Digital watermarking. San Francisco. Morgan Kaufmann.

[11] Caldelli, R., Filippini, F. and Barni, M. 2006. Joint near- lossless compression and watermarking of still images for authentication and tamper localization. Signal Processing: Image Communication. 21:890–903.

[12] Chang, Chin-Chen, Yi-Hsuan Fan, and Wei-Liang Tai. 2008. Four-scanning attack on hierarchical digital watermarking method for image tamper detection and recovery. Pattern Recognition. 41(2): 654-661.

[13] Zain, J. M. and Fauzi, A. R. M. 2006. Medical Image Watermarking with Tamper Detection and Recovery. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3270-3273.

[14] Sahu, A. K., & Sahu, M. (2020). Digital Image Steganography and steganalysis: A journey of the past three decades. Open Computer Science, 10(1), 296–342. https://doi.org/10.1515/comp-2020-0136.

[15] Sreenivas, K., & Kamkshi Prasad, V. (2017). Fragile watermarking schemes for image authentication: A survey. International Journal of Machine Learning and Cybernetics, 9(7), 1193–1218. https://doi.org/10.1007/s13042-017-0641-4.

[16] Prasad, S., & Pal, A. K. (2020). Hamming code and logistic-map based pixel-level active forgery detection scheme using fragile watermarking. Multimedia Tools and Applications, 79(29-30), 20897–20928. https://doi.org/10.1007/s11042-020-08715-x.

[17] Rakhmawati, L., Suwadi, S., & Wirawan, W. (2020). Blind robust and self-embedding fragile image watermarking for image authentication and copyright protection with Recovery Capability. International Journal of Intelligent Engineering and Systems, 13(5), 197– 210. https://doi.org/10.22266/ijies2020.1031.18.

[18] Hurrah, N. N., Parah, S. A., Loan, N. A., Sheikh, J. A., Elhoseny, M., & Muhammad, K. (2019). Dual watermarking framework for privacy protection and content authentication of multimedia. Future Generation Computer Systems, 94, 654–673. https://doi.org/10.1016/j.future. 2018.12.036.

[19] Hisham, S. I., Muhammad, A. N., Badshah, G., Johari, N. H., & Mohamad Zain, J. (2016). Numbering with spiral patterns to prove authenticity and integrity in medical images. Pattern Analysis and Applications, 20(4), 1129–1144. https://doi.org/10.1007/s10044- 016-0552-0.

[20] Canales, K. (2021, February 9). Twitter surpassed 192 million daily active users in Q4 as the social media company grappled with criticism of its role in the spread of election misinformation. Business Insider.

Retrieved June 21, 2022, from https://www.businessinsider.com/twitter-earnings-q4- revenue-eps-new-users-2021-2.

[21] Schneider, G.M. & Gersting, J.L. 2004. Invitation to computer science. Course Technology.

[22] Lossy vs lossless compression - keycdn support. KeyCDN. (n.d.). Retrieved July 29, 2022, from https://www.keycdn.com/support/lossy-vs-lossless.

[23] Patvardhan, C., Kumar, P., & Vasantha Lakshmi, C. (2017). Effective color image watermarking scheme using ycbcr color space and QR code. Multimedia Tools and Applications, 77(10), 12655–12677.

[24] Prinkle Rani, & Jyoti Rani. (2015). Copy-move forgery attack detection in digital images. International Journal of Engineering Research And, V4(06). https://doi.org/10.17577/ijertv4is061110.

[25] Hassan, A., & Sharma, V. K. (2021). Texture based image splicing forgery recognition using a passive approach. International Journal of Integrated Engineering, 13(4). https://doi.org/10.30880/ijie.2021. 13.04.010.

[26] Aditi. (2020, September 16). Classical art memes: A visual analysis. openclosemag. Retrieved June 21, 2022, from https://www. openclosemag.com/post/classical-art-memes-a- visual analysis.

[27] Ross, A. S., & Rivers, D. J. (2019). Internet memes, media frames, and the conflicting logics of climate change discourse. Environmental Communication, 13(7), 975–994. https://doi.org/10.1080/17524032. 2018 .1560347.

[28] Gaaed, M., & Tahar, M. (2018). Digital Image Watermarking based on LSB techniques: A comparative study. International Journal of Computer Applications, 181(26), 30–36. https://doi.org/10.5120 /ijca2018918105.

[29] Zain, J. M. and Fauzi, A. R. M. 2006. Medical Image Watermarking with Tamper Detection and Recovery. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3270-3273.

[30] Lin, C. and Chang, S. 2001. A Robust image authentication method distinguishing JPEG compression from malicious manipulation. IEEE Transactions on Circuits and Systems for Video Technology. 11(2): 153-168.

[31] Ulutas, G., &; Muzaffer, G. (2016). A new copy moves forgery detection method resistant to object removal with uniform background forgery. Mathematical Problems in Engineering,2016, 1–19. https://doi.org/10.1155/2016/321516.

[32] El-Alfy, E.-S. M., & Qureshi, M. A. (2014). Combining spatial and DCT based Markov features for enhanced blind detection of image splicing. Pattern Analysis and Applications, 18(3), 713–723. https://doi.org/10.1007/s10044-014-0396-4.

[33] Gaaed, M., & Tahar, M. (2018). Digital Image Watermarking based on LSB techniques: A comparative study. International Journal of Computer Applications, 181(26), 30–36. https://doi.org/10.5120/ijca 2018918105.

[34] Kumar, S., & Dutta, A. (2016). A novel spatial domain technique for digital image watermarking using block entropy. 2016 International Conference on Recent Trends in Information Technology (ICRTIT). https://doi.org/10.1109/icrtit.2016.7569530.

[35] Tao, H., Chongmin, L., Mohamad Zain, J., & Abdalla, A. N. (2014). Robust image watermarking theories and techniques: A Review. Journal of Applied Research and Technology, 12(1), 122–138. https://doi.org/10.1016/s1665-6423(14)71612-8.

[36] Fridrich, J., Goljan, M. and Du, R. (2001) Lossless data embedding – new paradigm in digital watermarking, EURASIP Journal of Applied Signal Processing, Vol.2, Pp. 185-196.

[37] Q. Li and N. Memon, “Security Models of Digital Watermarking," in Multimedia Content Analysis and Mining. vol. 4577, N. Sebe, Y. Liu, Y. Zhuang, and T. Huang, Eds., ed: Springer Berlin / Heidelberg, 2007, pp. 60-64.

[38] Shahreza, M. S. 2005. An Improved Method for Steganography on Mobile Phones.

[39] Zhang, H., Wang, C., & Zhou, X. (2017). Fragile Watermarking for image authentication using the characteristic of SVD. Algorithms, 10(1), 27. https://doi.org/10.3390/a1001002.

Computational Study of Quantum Coherence from Classical Nonlinear Compton Scattering with Strong Fields

Huber Nieto-Chaupis
Universidad Autónoma del Perú
Panamericana Sur Km. 16.3 Villa el Salvador
Lima Perú

***Abstract*—From the covariant formulation of radiation intensity of Hartemann-Kerman model entirely constructed in the classical electrodynamics scenario, a formulation of coherent states has been obtained in an explicit manner represented by the infinite sum of integer-order Bessel functions. Both linear and nonlinear Compton scattering are included, suggesting that Compton processes can be perceived as coherent states of light-matter interaction.**

***Keywords*—Quantum coherence; bessel; compton scattering**

I. Introduction

Compton scattering is seen as a "golden processes" inside Quantum Electrodynamics [1]. This has played in the understanding role in the understanding of quantum mechanics of light-matter interactions. In fact it is pure quantum effect in the which the electron absorbs one single photon and emits one photon with different kinematics than the first one. In a full quantum theory, the Lagrangian of interaction can be written as:

$$\mathcal{L}_{\rm INT} = -ie\int dx^4 \bar{\Psi}\gamma_\mu A^\mu \Psi, \tag{1}$$

with $\bar{\Psi}$ and Ψ the final and initial states, while γ the 4×4 matrices, and A^μ the 4-vector potential that satisfies the Lorentz's gauge $\partial_\mu A^\mu$ = 0. In a nutshell the incorporation of a propagator in Eq. (1) and the subsequent operations yields commonly the well-known Feynman's diagrams [2]. Compton scattering was also boarded at the scenario of strong electromagnetic fields. In this case the states $\bar{\Psi}$ and Ψ are solutions of Volkov and obey the equation of Dirac with a external field. In the scenario of high regime where the incoming electromagnetic field has a high density of photons it is usual to define the intensity parameter:

$$\xi^2 = \frac{e^2\mathbf{A}^2}{m}. \tag{2}$$

introduced by I.I. Goldman [3] who derived the energy of emitted photon given by:

$$\omega' = \frac{2nE\omega}{E(1+{\rm Cos}\theta) + [2n\omega + \frac{m^2(1+\xi^2)}{E}(1-{\rm Cos}\theta)]}, \tag{3}$$

where the product $n\omega$ is denoting the absorption of n photons. From this $n = 1$ the Klein-Nishina formula is restored. The integer n is linked to the nonlinear processes at the which the electron can absorb various photons simultaneously. These non-linearities can be compacted in the language of Quantum Electrodynamics (QED in short) as a Dirac-Delta function:

$$\delta(E_{\rm I} + n\omega - E_{\rm F} - m\omega'), \tag{4}$$

that indicates the conservation of energy with m an integer number. This non-linearity is theoretically obtained in the emission and absorption of various laser photons by Reiss [4] and Ritus [5] whom have derived quantization of laser in a semi-classical arena with the laser modeled through an circularly polarized infinite wave. This was also seen at [6] and the works of Eberly [7]. A noteworthy attention was paid at the 90s because the prospective construction of a photon-photon collider [8] and the potential apparition of non-linearities as corroborated at the experiments observed at SLAC [9] where nonlinear Compton was observed with strong lasers supporting the fact that these processes can be well modeled by an infinite classical monochromatic wave. In photon collisions one expects that the Compton backscattering can create new particles in according to the reactions:

$$\gamma + \gamma \Rightarrow \tilde{\ell}^+ + \tilde{\ell}^- + \sum_q^Q \Xi_q$$

with the production of Q particles, was predicted inside the framework of new physics of elementary particles [10]. In 1996 nonlinear Compton backscattering have been obtained in an entire arena of classical electrodynamics by Hartemann and Kerman [11] (HK model in short) from the intensity of radiation $\frac{dI(\omega)}{d\Omega} =$

$$\frac{\omega^2}{4\pi^2c^2}\left|\int dt\int d^3x\mathbf{n}\times[\mathbf{n}\times\mathbf{J}(\mathbf{x},t)]e^{i\omega[t-\frac{\mathbf{n}\cdot\mathbf{x}}{c}]}\right|^2. \tag{5}$$

Here, it was shown that Compton scattering governs the low intensity whereas in super-strong fields the nonlinear Compton scattering emerges as the apparition of high harmonics that are interpreted as emission of photons with different frequencies. Based at all this background where quantum effects can be retrieved from classical formalisms, this paper tries to derive the quantum coherence from the HK theory. Inspired at the theory of Glauber [12][13], coherent states

proportional to Bessel functions are derived. In second section Compton processes are derived from classical electrodynamics. In third section the quantum coherence is derived. Finally at last section the conclusion of paper is presented.

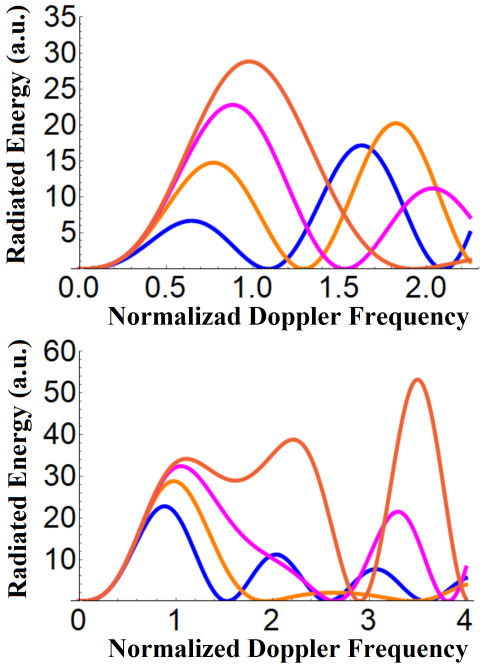


Fig. 1. Classical Distributions of Radiated Energy from Eq. 7 Once the Bessel Expansion as Done at Eq. 14 are Plotted. Up: q Runs from 2 to 5. Down: q from 4 to 7. It should be Noted the Peaks at χ=1.0, Denoting the Simple Classical Compton Scattering.

II. Classical Nonlinear and Linear Compton Scattering

The HK model was developed under a covariant framework in the sense that $\phi = k_\mu \cdot x^\mu$ yielding the well-known radiation intensity of a single electron in an external super intense polarized laser. The radiation intensity is depending on the Doppler-shifted frequency χ. As seen in HK model, Compton and nonlinear Compton scattering was obtained for different values of pulse width. Therefore, one can arrive to the fundamental equation of HK model that can be written down as the distribution of energy radiated per solid angle and frequency and with the definition:

$$\lambda = \frac{e^2}{4\pi^2} u_0^2 \chi^2, \tag{6}$$

then the fundamental equation of covariant HK model can be written as:

$$\frac{d^2I}{d\omega d\Omega} = \lambda \left| \int_{-\infty}^{+\infty} A_x(\phi) \exp\left\{ i\chi \left[\phi + \int_{-\infty}^{\phi} \mathbf{A}^2(\psi) d\psi \right] d\phi \right\} \right|^2. \tag{7}$$

Clearly one can appeal to different mathematical approaches to extract the quantum mechanics (if any) of Eq. 7 in different ways. At [14], from the HK model the argument of Dirac delta functions have been obtained as well as interpreted as the absorption and emission of photons even when the external field was an infinite wave as commonly expected from QED. Although of course the applicability of advanced mathematical methodologies cannot guarantee not any kind of quantization of external field, a suitable methodology turns out to be the usage of the Fourier expansion. In fact, consider the identity based in the series of Fourier-Bessel so that from the exponential of Eq.7 one gets:

$$\mathrm{Exp}\left\{ i\chi \left[\phi + \int_{-\infty}^{\phi} \mathbf{A}^2(\psi) d\psi \right] \right\} = \sum_{\infty}^{\infty} J_q(\chi) \mathrm{Exp}[iq\theta], \tag{8}$$

with the usage of the crude approximation:

$$\sin\theta = \left[\phi + \int_{-\infty}^{\phi} \mathbf{A}^2(\psi) d\psi \right] \tag{9}$$

$$\Rightarrow \theta = \sin^{-1} \left[\phi + \int_{-\infty}^{\phi} \mathbf{A}^2(\psi) d\psi \right], \tag{10}$$

that to some extent θ can be seen as a phase. In this manner by putting Eq. 8 and Eq. 9 into Eq. 7 one can see that the Bessel functions can be out of the integration. With this Eq. 7 can be written in a more transparent manner as:

$$\frac{d^2I}{d\omega d\Omega} = \lambda \left| \sum_{-\infty}^{+\infty} J_q(\chi) \right|^2 \left| \int_{-\infty}^{+\infty} d\phi A_x(\phi) \mathrm{Exp}\left(iq\mathbf{sin}^{-1}[\theta(\phi)] \right) \right|^2. \tag{11}$$

As expressed in the HK model, the external field is a super intense laser that is characterized by the width $\Delta\phi$ that is entirely an experimental input. Therefore one can define a function depending on ϕ in the sense that:

$$F(\Delta\phi) = \left| \int_{-\infty}^{+\infty} d\phi A_x(\phi) \mathrm{Exp}\left(iq\mathbf{sin}^{-1}[\theta(\phi)] \right) \right|^2. \tag{12}$$

It is because once the integration is done through the variable ϕ it yields only a pure dependence on the pulse's width $\Delta\phi$ then one can rewrite Eq.11 as:

$$\frac{d^2I}{d\omega d\Omega} = \lambda \left| \sum_{-\infty}^{+\infty} J_q(\chi) \right|^2 F(\Delta\phi). \tag{13}$$

Subsequently, one can arrive to a normalized backscattered spectrum that would depend on the Doppler-shifted frequency χ. On the other hand by knowing the input value for $\Delta\phi$ then $F(\Delta\phi)$ can opt a finite value, for instance ρ. When λ is written in an explicit manner from Eq. 6 and inserting it into Eq. 13 then the resulting radiation intensity can be written as:

$$\frac{4\pi^2}{e^2\mu_0^2\rho} \frac{d^2I}{d\omega d\Omega} = I(Q,\chi) = \chi^2 \left| \sum_{q}^{Q} J_q(\chi) \right|^2. \tag{14}$$

The way as it is written Eq. 14 allows to displayed it in a straightforward manner. In fact, in Fig. 1 the normalized backscattered spectrum is plotted for two scenarios. Here $\frac{4\pi^2}{e^2\mu_0^2\rho} \approx \xi$. For this exercise, the Up-panel displays various

curves of radiated classical energy. Here the sum ran from 2 to the 5th harmonic as given by:

$$I(Q,\chi) = 25\chi^2 \left| \sum_{q=2}^{Q} J_q(3.2\chi + 1) \right|^2 . \quad (15)$$

One can see there, the Grey color line is denoting the sum of all 4 orders, and it is peaked denoting the fact that still at the classical formulation, scattering Compton can be derived. At the Down-panel where the sum runs from 4 to 7, is exhibiting for instance the Grey liine, a deformed shape in contrast to Up-panel.

$$I(Q,\chi) = 25\chi^2 \left| \sum_{q=4}^{Q} J_q(3.2\chi + 1) \right|^2 . \quad (16)$$

In on the other side, the Grey line the sum of all four orders, is revealing that high orders might be distorting the peaked centered at χ=1. It is because the highest order are certainly connected to nonlinear Compton scattering. In fact, such deformation at the Grey line is due also to the contribution of more photons to the state of absorption by the electron at the strong electromagnetic field, so that the electron has much energy to emit. Of course, although it is argued in a fully classical scenario, the implementation of integer-order Bessel functions, allows to examine the radiated energy spectra from this perspective. Under the hypothesis that Eq. 14 is an element of an infinite sum then one can generalize it with the change $\xi = \frac{4\pi^2}{e^2\mu_0^2\rho}$ one can write below that:

$$1 + \xi \frac{d^2I}{d\omega d\Omega} \approx 1 + \frac{\chi^2}{2!} \left| \sum_{-\infty}^{+\infty} J_q(\chi) \right|^2 + ... + \frac{\chi^n}{n!} \left| \sum_{-\infty}^{+\infty} J_q(\chi) \right|^n \Rightarrow \xi \frac{d^2I}{d\omega d\Omega} = \mathrm{Exp}\left\{ \left| \chi \sum_{-\infty}^{+\infty} J_q(\chi) \right|^2 \right\}, \quad (17)$$

so that Eq. 9 with the hypothesis that $\chi^n = 0$ for $n \geq 3$ thus it can finally be written as:

$$\xi \frac{d^2I}{d\omega d\Omega} = \mathrm{Exp}\left\{ - \left| \chi \sum_{-\infty}^{+\infty} J_q(\chi) \right|^2 \right\}. \quad (18)$$

III. Derivation of Quantum Coherence

The mathematical structure of Eq. 10 allows to link it to the quantum mechanics territory in the sense that quantum coherence can be extracted. For this one should assume the following hypothesis:

$$\left| \chi \sum_{q=-\infty}^{Q=+\infty} J_q(\chi) \right|^2 = \frac{\alpha^2}{2}, \quad (19)$$

by which one arrives to:

$$\xi \frac{d^2I}{d\omega d\Omega} = \mathrm{Exp}\left(-\frac{|\alpha^2|}{2} \right), \quad (20)$$

that in the quantum scenario on gets:

$$|\langle 0|\alpha\rangle|^2 = \mathrm{Exp}\left(-\frac{|\alpha^2|}{2} \right), \quad (21)$$

with α the eigenvalue of the equation $\hat{\mathbf{a}}|\alpha\rangle = \alpha|\alpha\rangle$. From Eq. 20 and Eq. 21 one arrive to:

$$\xi \frac{d^2I}{d\omega d\Omega} = |\langle 0|\alpha\rangle|^2 = \mathrm{Exp}\left\{ - \left| \chi \sum_{-\infty}^{+\infty} J_q(\chi) \right|^2 \right\}, \quad (22)$$

so that one finds that the quantum mechanics amplitude can be written in terms of classical electrodynamics observables:

$$|\langle 0|\alpha\rangle| = \sqrt{\xi \frac{d^2I}{d\omega d\Omega}}, \quad (23)$$

and the eigen values of coherence can be expressed in terms of interger-order Bessel functions:

$$\alpha^2 = 2 \left| \chi \sum_{-\infty}^{+\infty} J_q(\chi) \right|^2 , \quad (24)$$

indicating that the values of coherence depend on the χ variable, the normalized Doppler-shifted frequency. It should be noted the relevance of orthogonal polynomial at the classical formulation of coherence [15][16][17][18] In other words, the coherence is directly linked to the frequencies of the emitted photons (or another observable as commonly done in quantum mechanics [19][20][21][22][23][24]). Indeed, the eigenvalues equation involving the annihilation operator and the states of coherence is written below as:

$$\hat{\mathbf{a}}|\alpha\rangle = \alpha|\alpha\rangle = \sqrt{2} \sum_{-\infty}^{+\infty} \chi J_q(\chi)|\alpha\rangle \quad (25)$$

In Fig. 2 (Up panel) square of coherence Eq. 19 as well as $|\langle 0|\alpha\rangle|^2$ have been plotted as function of normalized Doppler-shift frequency. Interestingly in left-side up two peaks for $2 < \chi < 10$ The one of interest (blue line Q=2) because one finds minor peaks $\chi = 3, 6$ and $\chi = 9$ as well as one can see a large peak at $\chi = 10$ for Q=10. In the right-side it is easy to note that all lines are centered at $\chi = 0$ indicating that the classical view the system has null energy to emit photons at the Compton range, however one can see minor peaks for $\chi > 4$. In (Down panel) the square of amplitude $|\langle 0|\alpha\rangle|^2$ is plotted.Here one can see that the orange line Q=10, appears to be deeply degraded. In the contrary case. the blue line denoting Q=2 exhibits high values above 50%. Thus one can see that while the lowest values of Q exhibit high values, thus one can see that the orders of Bessel function dictated by:

$$x^2 \frac{d^2y}{dx^2} + x\frac{dy}{dx} + (x^2 - q^2)y = 0 \quad (26)$$

might be relevant for $n = 0$ yielding (after of dividing over x^2) one arrives to:

$$\frac{d^2y}{dx^2} + \frac{1}{x}\frac{dy}{dx} + y = 0 \quad (27)$$

that might be strongly related to the high values of $|\langle 0|\alpha\rangle|^2$

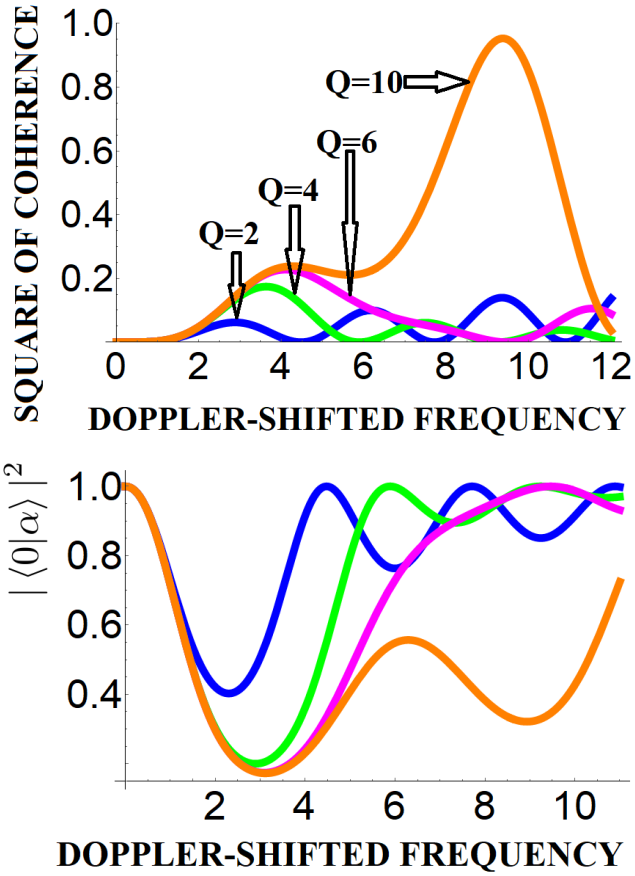


Fig. 2. (Up-Panel) Plotting of Eq. 19 Normalized to 1 for up to 4 Values of Integer Q. One can See that the Normalized Coherence Plotted as it Square Acquires a Similar form as the Radiation Intensities Plotted at Fig. 1. (Down-Panel) Plotting of Eq. 22 with Colors Same as Left-Panel Indicating that not any Peak at $\chi = 1$. Plots of Up and Down Panels were done with the Package [25].

A. Identifying the Doppler-Shifted Frequency Orthogonal Basis

The mathematical structure derived at Eq. 25 might be suggesting the existence of an orthogonal basis whose dependence would be given at the variable χ (see for example [26][27][28][29][30]). In fully accordance to the quantum mechanics formalism it is feasible to write down the completeness relationship for the normalized Doppler-shifted frequency χ as follows:

$$\int d\chi \, |\chi\rangle \langle\chi| = \mathbb{I}, \tag{28}$$

and in the other hand one has also that: $\int_{-\infty}^{\infty} d\alpha \, |\alpha\rangle \langle\alpha| = \mathbb{I}$ in conjunction to a discrete basis: $\sum_q |q\rangle \langle q| = \mathbb{I}$. In this way one can combine all these completeness relationships to arrive to:

$$\mathbb{I} \otimes \mathbb{I} = \sum_q \int d\chi \, \langle\chi|q\rangle \, |\chi\rangle \langle q| \, . \tag{29}$$

Now, one can include the coherent states also at the chain of multiplication of unitary operators in the form:

$$\begin{aligned} \mathbb{I} \otimes \mathbb{I} \otimes \mathbb{I} &= \int_{-\infty}^{\infty} d\alpha \, |\alpha\rangle \langle\alpha| \sum_q \int d\chi \, \langle\chi|q\rangle \, |\chi\rangle \langle q| \\ &= \sum_q \int d\chi \int_{-\infty}^{\infty} d\alpha \, |\alpha\rangle \langle\chi|q\rangle \langle\alpha|\chi\rangle \langle q| \end{aligned}$$

and the multiplication by the ket $|q\rangle$ in both sides one arrives to:

$$\begin{aligned} \mathbb{I} \otimes \mathbb{I} \otimes \mathbb{I} \, |q\rangle &= \sum_q \int d\chi \int_{-\infty}^{\infty} d\alpha \, |\alpha\rangle \langle\chi|q\rangle \langle\alpha|\chi\rangle \langle q|q\rangle \\ &= \sum_q \int d\alpha \, |\alpha\rangle \langle\alpha|\chi\rangle \int_{-\infty}^{\infty} d\chi \, \langle\chi|q\rangle \end{aligned}$$

where $\langle q|q\rangle = 1$. Indeed with the assumption:

$$\langle\alpha|\chi\rangle = \left(\frac{\alpha}{\chi}\right) \tag{30}$$

$$\langle\chi|q\rangle = \left(\chi \frac{d(\chi J_q(\chi))}{d\chi}\right) \tag{31}$$

then one gets from Eq. 27 in a straightforward manner the integration over χ:

$$\int_{-\infty}^{\infty} d\chi \frac{d(\chi J_q(\chi))}{d\chi} = \chi J_q(\chi) \tag{32}$$

so that the integral over α is trivial:

$$\int_0^{\sqrt{2\sqrt{2}}} d\alpha\alpha = \sqrt{2} \tag{33}$$

and by putting these integrations into Eq.27 then one can see that Eq. 25 is restored:

$$\mathbb{I} \otimes \mathbb{I} \otimes \mathbb{I} \, |q\rangle = |q\rangle = \sqrt{2} \sum_{q=-\infty}^{+\infty} \chi J_q(\chi) \, |\alpha\rangle \tag{34}$$

and multiplying Eq. 34 by the $\langle q|$ in both sides and with the definition of the polynomial $\langle q|\alpha\rangle \Rightarrow \langle\alpha|q\rangle = \left(\alpha \frac{d(\alpha J_q(\alpha))}{d\alpha}\right)$ derived from Eq. 31, then one can arrive to:

$$\sqrt{2} \sum_{q=-\infty}^{+\infty} \chi J_q(\chi) \left(\alpha \frac{d(\alpha J_q(\alpha))}{d\alpha}\right) = \mathbb{I} \tag{35}$$

Therefore, the derivative can be carry out:

$$\sum_{q=-\infty}^{+\infty} \left[\alpha\chi J_q^2(\alpha) + \chi\alpha^2 J_q(\chi) \frac{dJ_q(\alpha)}{d\alpha}\right] = \frac{1}{\sqrt{2}} \tag{36}$$

IV. Conclusion

In this paper, the quantum coherence has been extracted from the backscattered radiation intensity obtained in classical electrodynamics. While the HK model has been used, the results of this paper confirms that in the super intense regime the classical picture can restore quantum effects in particular the coherence of emitted radiation that to some extent is encompassing with requirements of advanced experiments that require backscattered radiation at the GeV energies to create unseen states of matter. Because these results can be understood as preliminary, in a next work the derivations from the HK model and its validation with current theories of quantum optics [31][32] shall be done.

References

[1] Feynman, Richard P. (1949). Space-Time Approach to Quantum Electrodynamic. Physical Review. 76 (6): 769–789.

[2] James D Bjorken; Sidney D Drell, New York : McGraw-Hill, 1965.

[3] I. I. Goldman, Intensity effects in compton scattering, Physics Letters Volume 8, Issue 2, 15 January 1964, Pages 103-106.

[4] Howard R. Reiss and Joseph H. Eberly, Green's Function in Intense-Field Electrodynamics, Phys. Rev. 151, 1058 (1966) - Published 25 November 1966.

[5] V. I. Ritus, "Quantum effects of the interaction of elementary particles with an intense electromagnetic field", J. Russ., Laser Res. 6 (1985), no. 5, 497.

[6] A. I. Nikishov and V. I. Ritus, Quantum processes in the field of a plane electromagnetic wave and in a constant field I, Sov. Phys. JETP 19 (1964), no. 2, 529–541.

[7] J. H. Eberly and H. R. Reiss, Electron Self-Energy in Intense Plane-Wave Field, Phys. Rev. 145, 1035 (1966).

[8] Florian Bechtel and $et.al$, Studies for a photon collider at the ILC, NIM-A, Volume 564, Issue 1, 1 August 2006, Pages 243-261.

[9] C. Bula and $et.al$, Observation of Nonlinear Effects in Compton Scattering, Phys. Rev. Lett. 76, 3116 – Published 22 April 1996.

[10] Nieto-Chaupis, Huber, Study of scalar leptons at the TESLA photon collider, PhD thesis, 2008, edoc-Server Open-Access-Publikationsserver der Humboldt-Universitat zu Berlin.

[11] F. V. Hartemann and A. K. Kerman, Classical Theory of Nonlinear Compton Scattering, Phys. Rev. Lett. 76, 624 (1996) - Published 22 January 1996.

[12] Roy J. Glauber, Photon Correlations, Phys. Rev. Lett. 10, 84 (1963) - Published 1 February 1963.

[13] Roy J. Glauber, The Quantum Theory of Optical Coherence, Phys. Rev. 130, 2529 (1963) - Published 15 June 1963.

[14] Huber Nieto-Chaupis, Quantum Effects without Quantum Fields: Feynman's Amplitudes in Classical Electrodynamics, 2019 IEEE 2nd British and Irish Conference on Optics and Photonics (BICOP).

[15] Kimmo Saastamoinen, Jari Turunen, Pasi Vahimaa, and Ari T. Friberg, Spectrally partially coherent propagation-invariant fields, Phys. Rev. A 80, 053804 (2009) - Published 3 November 2009.

[16] Alessandro Averchi, Daniele Faccio, Ricardo Berlasso, Miroslav Kolesik, Jerome V. Moloney, Arnaud Couairon, and Paolo Di Trapani, Phase matching with pulsed Bessel beams for high-order harmonic generation, Phys. Rev. A 77, 021802(R) (2008) - Published 20 February 2008.

[17] S. E. Harris and A. V. Sokolov, Subfemtosecond Pulse Generation by Molecular Modulation, Phys. Rev. Lett. 81, 2894 (1998) - Published 5 October 1998.

[18] A. Nazarkin, G. Korn, M. Wittmann, and T. Elsaesser, Generation of Multiple Phase-Locked Stokes and Anti-Stokes Components in an Impulsively Excited Raman Medium, Phys. Rev. Lett. 83, 2560 (1999) - Published 27 September 1999.

[19] Yuri Dakhnovskii, Dynamics of a two-level system with Ohmic dissipation in a time-dependent field, Phys. Rev. B 49, 4649 (1994) - Published 15 February 1994.

[20] Fam Le Kien, Anil K. Patnaik, and K. Hakuta, Multiorder coherent Raman scattering of a quantum probe field, Phys. Rev. A 68, 063803 (2003) - Published 2 December 2003.

[21] T. Mizushima and K. Machida, Splitting and oscillation of Majorana zero modes in the p-wave BCS-BEC evolution with plural vortices, Phys. Rev. A 82, 023624 (2010) - Published 31 August 2010.

[22] Yukio Shibata, Shigeru Hasebe, Kimihiro Ishi, Shuichi Ono, Mikihiko Ikezawa, Toshiharu Nakazato, Masayuki Oyamada, Shigekazu Urasawa, Toshiharu Takahashi, Tomochika Matsuyama, Katsuhei Kobayashi, and Yoshiaki Fujita, Coherent Smith-Purcell radiation in the millimeter-wave region from a short-bunch beam of relativistic electrons, Phys. Rev. E 57, 1061 (1998) - Published 1 January 1998.

[23] Chao Hang and Guoxiang Huang, Ultraslow helical optical bullets and their acceleration in magneto-optically controlled coherent atomic media, Phys. Rev. A 87, 053809 (2013) - Published 8 May 2013.

[24] Y. F. Chen, Y. C. Lin, W. Z. Zhuang, H. C. Liang, K. W. Su, and K. F. Huang, Generation of large orbital angular momentum from superposed Bessel beams corresponding to resonant geometric modes, Phys. Rev. A 85, 043833 (2012) - Published 20 April 2012.

[25] https://www.wolfram.com/mathematica/ .

[26] Lan Zhou, S. Yang, Yu-xi Liu, C. P. Sun, and Franco Nori, Quantum Zeno switch for single-photon coherent transport, Phys. Rev. A 80, 062109 (2009) - Published 11 December 2009.

[27] Peter S. M. Townsend and Alex W. Chin, Disentangling theorem and scattering functions for long-range coherent tunneling, Phys. Rev. A 99, 012112 (2019) - Published 14 January 2019.

[28] Mathias Wagner, Strongly Driven Quantum Wells: An Analytical Solution to the Time-Dependent Schrödinger Equation, Phys. Rev. Lett. 76, 4010 (1996) - Published 20 May 1996.

[29] Taiwang Cheng, Xiaofeng Li, Shuyan Ao, Ling-An Wu, and Panming Fu, Frequency-domain interpretation of the plateaus in laser-assisted recombination and high-order harmonic generation, Phys. Rev. A 68, 033411 (2003) - Published 25 September 2003.

[30] Mathias Wagner, Photon-assisted transmission through an oscillating quantum well: A transfer-matrix approach to coherent destruction of tunneling, Phys. Rev. A 51, 798 (1995) - Published 1 January 1995.

[31] D. J. Daniel and G. J. Milburn, Destruction of quantum coherence in a nonlinear oscillator via attenuation and amplification, Phys. Rev. A 39, 4628 (1989) - Published 1 May 1989.

[32] U. Rathe, M. Fleischhauer, Shi-Yao Zhu, T. W. Hänsch, and M. O. Scully, Nonlinear theory of index enhancement via quantum coherence and interference, Phys. Rev. A 47, 4994 (1993) - Published 1 June 1993.

Cybersecurity Risk Assessment: Modeling Factors Associated with Higher Education Institutions

Rachel Ganesen[1], Asmidar Abu Bakar[2], Ramona Ramli[3], Fiza Abdul Rahim[4], Md Nabil Ahmad Zawawi[5]
College of Graduate Studies, Universiti Tenaga Nasional, Malaysia[1]
College of Computing and Informatics, Universiti Tenaga Nasional, Malaysia[2, 3, 5]
Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Malaysia[4]
Institute of Informatics and Computing Energy, Universiti Tenaga Nasional, Malaysia[2, 3, 4, 5]

***Abstract*—Most universities rely heavily on Information Technology (IT) to process their information and support their vision and mission. This rapid advancement in internet technology leads to increased cyberattacks in Higher Education Institutions (HEIs). To secure their infrastructure from cyberattacks, they must implement the best cybersecurity risk management approach, which involves technological and education-based solutions, to safeguard their environment. However, the main challenges in existing cybersecurity risk management approaches are limited knowledge of how organizations can determine or minimize the significance of risks. As a result, this research seeks to advance understanding to establish a risk assessment model for universities to measure and evaluate the risk in HEIs. The proposed model is based on theoretical aspects that we organized as follows: First, we review the existing cybersecurity frameworks to identify the suitability and limitation of each model. Next, we review current works on cybersecurity risk assessment in HEIs to evaluate the proposed risk assessment approaches, scope and steps. Based on the information gathered, we developed a risk assessment model. Finally, we conclude the study with directions for future research. The result presented from this study may give an insig1ht for HEIs staff to analyze what is to be assessed, how to measure the severity of the risk, and determine the level of risk acceptance, improving their decision-making on risk management.**

***Keywords*—Cyber security; risk assessment; university**

I. Introduction

Higher education institutions (HEIs) are prime targets for cybercriminals because their networks hold sensitive personal information about students, including their academic and financial data. Several education organizations and institutions have been victims of cyberattacks [1]. Cybercriminals in Asia exploit flaws in IT systems that support schools and universities in carrying out various attacks. Even before the pandemic, a massive data breach that had reportedly hit a prominent Malaysian university resulted in the personal data of over one million people being leaked online [2].

During the COVID-19 pandemic, every industry faces significant change and ongoing challenges. Like many other industries, the higher education sector has been overturned by the COVID-19 pandemic. In place of classroom instruction, many students are learning virtually and remotely. While the shift to remote education may have helped the governments better contain the spread of COVID-19, it is also added a layer of cybersecurity risks that higher education institutions (HEIs) are forced to confront.

When the pandemic forced HEIs to use online platforms to conduct classes and evaluates students, it created a new entry point for cybercriminals to target due to the vulnerabilities in online platforms. These platforms include video chat programs like Zoom and Microsoft Teams and curriculum, technology, and services providers. According to Malwarebytes, the education sector is the top target for Trojan malware [3]. Kaspersky discovered 356,000 malicious files while investigating infected online textbooks, including 233,000 malware-infected essays and 123,000 malware-infected books [4]. A recent Kaspersky study showed that the number of users exposed to various threats using common online learning sites as a lure reached 270,171 in January 2021, up 60% from the first half of 2020 [5]. The rapid development of internet technologies and online platforms among students has led to increased cyberattacks in HEIs.

Since new and more advanced threats arise at an unprecedented pace, it is evident that HEIs are at risk of potentially disastrous security incidents if adequate security measures and workforce preparation initiatives are not implemented. Representatives from every campus department, such as administration, facilities, communications, and IT, must work together to analyze potential risks and create policies to address them [6]. To secure their infrastructure from cyberattacks, HEIs must implement the best cybersecurity risk assessment approach, which involves technological and education-based solutions, to safeguard their HEI environment.

Risk assessment provides organizations with an accurate evaluation of the risks to their assets. It can help them prioritize and develop a comprehensive strategy to reduce risks [7]. As highlighted by Panchal [8], many institutions have limited or no visibility of their IT risk exposure. Furthermore, available resources are not utilized effectively to manage the risks. The primary concerns in current risk assessment methodologies are how HEIs can estimate the significance of risks and develop resolution capabilities to deal with or minimize the risks. [9].

Therefore, this study aims to establish a cybersecurity risk assessment model for HEIs. The proposed model is based on theoretical aspects that we organized as follows: First, we review the existing cybersecurity frameworks to identify the suitability and limitation of each model. Next, we review current works on cybersecurity risk assessment in HEIs to

evaluate the risk assessment approaches, risk metrics and steps proposed. We developed a risk assessment model combining ISO 27005 and NIST SP 800-30 framework based on the information gathered.

II. Review of Current Cybersecurity Risk Assessment Literature

A. Cybersecurity Risk Assessment Frameworks

Risk assessment is an important methodology for cybersecurity that employs techniques to assist organizations in dealing with uncertain events [10]. It is a tool for assessing factors that contributes to a failure or loss that hinders the success of a project or business. Various risk assessment models are available, some of which are qualitative while others are quantitative, with a common goal of estimating the overall risk value.

The Software Engineering Institute developed OCTAVE (Operationally Critical Threat, Asset and Vulnerability Evaluation) at Carnegie Mellon University to help the U.S. Department of Defense (DoD) address its security risks and challenges [11]. OCTAVE has two variants: OCTAVE-S and OCTAVE Allegro [12]. OCTAVE-S has fewer processes, adhering to the overall OCTAVE philosophy and thus simplifying application for small businesses. OCTAVE Allegro is a later variant focused on protecting information-based critical assets. The OCTAVE framework is workshop-oriented, requiring knowledge from three levels: senior management, operational area management, and staff. Many risk assessment practitioners agree that the detail level and complexity of the OCTAVE assessment approach have made it hard to adopt on a wide scale [13].

Facilitated Risk Analysis Process (FRAP) is a method where information security provision is considered as part of the risk management process. The main objective of the Facilitated Risk Analysis Process (FRAP) was to develop an efficient and disciplined process to ensure that information-related risks to business operations are considered and documented [14]. Table I shows how each risk analysis procedure is separated into three distinct sessions.

However, this model requires expert communications and internal managers' participation to collect data, making the process more time-consuming. Besides that, this framework is designed to analyze business and not comply with security requirements.

Another prominent framework is ISO 27005, the international standard that guides information security risk management processes that are needed for the implementation of an effective information security management system (ISMS) [13]. The stages of risk assessment consist of context establishment, risk identification, risk analysis, risk evaluation, and risk management [15]. ISO 27005 provides good examples of a threat catalogue, vulnerabilities, and various computation and plotting techniques for rating risk. However, the limitation of this framework is that it focuses on objectives, guidance, and concept but does not provide any criteria, scoring, or decision matrices.

National Institute of Standards and Technology (NIST) published the latest version of the Cybersecurity Framework. This framework categorizes cybersecurity practices in five domains: Identify, Protect, Detect, Respond and Recover. As for the NIST method, the risk assessment process is refined into nine steps. Each step has a clear goal and all the possible approaches to accomplish the goal, which alleviates the bias brought by merely depending on participants' or security evaluator's knowledge [16].

Table II summarizes the differences between all four frameworks. Each framework has been categorized based on five criteria: phases, data collection method, approach and complexity. The OCTAVE framework phases focus more on assets while ISO 27000 and NIST focus on data security. The FRAP framework focuses more on business analysis than security assessment. The data collection method for OCTAVE and FRAP is largely dependent on the participants' knowledge which can be time-consuming. Meanwhile, NIST's and ISO 27005 framework data collection method is not limited to participants' knowledge but includes conclusions and discoveries mentioned in other related documentation.

In terms of approach, the OCTAVE framework is based on methodology and has an implementation guide. The FRAP framework is based on guidelines and participants' decisions. Meanwhile, ISO 27005 focuses on objectives, guidelines, and concepts and does not really provide criteria, scoring, or decision matrices. The NIST framework enumerates all the possible approaches to process the data and has a specific target to facilitate the procedure.

The complexity of each framework is defined by the time consumed to process and gather the data, and it can be categorized as high, medium and low. High complexity requires more participation in data collection and more time to process the data. Medium complexity is when it requires an average number of participants in data collection and an average time to process the data. In contrast, low complexity is when fewer people are required for data collection and less time is required to process the data. As a result, NIST SP 800-30 and ISO 27005 frameworks provide the most complete and scientific approach among all the methods.

TABLE I. Risk Analysis Procedure in FRAP

FRAP Session	Description
PRE FRAP	It takes about an hour and involves the business manager, project lead and facilitator. The project outcome depends on five key components: scopes statement, visual mode, FRAP team, meeting mechanics and agreement on definitions.
FRAP SESSION	It takes between 7 and 15 hours to complete and includes 15 people in the organization. The second session is to access threats with the existing control place. It has three phases: risk analysis, safeguard implementation and security assessment.
POST FRAP	It takes about an hour with the same attendees. The deliverables for this meeting include a summary of threats and existing controls, as well as a final report.

TABLE II. COMPARISON OF FOUR CYBERSECURITY RISK ASSESSMENT FRAMEWORKS

Framework	Phases	Data Collection Method	Approach	Complexity
OCTAVE	1. Development of a profile of threats related to the asset 2. Identification of vulnerabilities 3. Development of security strategies and plans	Requires knowledge from all three levels: senior management, operational level and steps but does not imply third-party experts	Method based	High
ISO 27005	1. Context of risk establishment 2. Risk Identification 3. Risk Analysis 4. Risk Evaluation	Requires knowledge from internal managers	Guidelines	Low
FRAP	1. Pre frap meeting 2. FRAP session 3. Post FRAP process	Requires knowledge from internal managers and experts	Guidelines	Medium
NIST SP800-30	1. System characterization 2. Threat identification 3. Vulnerability identification 4. Control analysis 5. Likelihood determination 6. Impact analysis 7. Risk determination 8. Control recommendations 9. Results documentation	Non-government organizations	Guidelines	Medium

B. Related Works on Cybersecurity Risk Assessment in HEIs

Jufri et al. [17] conducted a risk assessment on the Academic Information System asset on OCTAVE Allegro and ISO framework. This research focuses on the Academic Information System in Langlangbuana University that functions to protect its critical assets. The process of risk assessment is conducted based on the OCTAVE framework. The implementation of security control is based on ISO 27002.

Similarly, Chanchala Joshi [18] has also proposed a quantitative information risk assessment model based on the OCTAVE framework for the university computing environment. The proposed model quantitatively measures security risks by identifying threats and information processes within university network configuration. The first phase focuses on knowing weak points. The next phase concentrates on understanding which areas have the highest risks. The last phase pivots with creating an actionable remediation plan over the university environment's unique factor and finally generate powerful reporting to track recursive risk measurement activities. The major drawbacks of OCTAVE are its complexity and that it does not allow organizations to quantitatively model risk. In order to improve the security organization system, some standard principles are required.

Meanwhile, Hom et al. [19] and Suroso et al. [20] proposed a risk assessment model to identify, analyze and manage the risk of academic information systems in higher education using the OCTAVE Allegro method. The risk assessment was conducted based on four stages, where first they establish drivers, profile assets, identify threats and mitigate risks. This approach differs from the OCTAVE approach because OCTAVE Allegro focuses on information assets within the context of how they are used, where they are stored, transported and processed, and how they are affected by the threat, vulnerability, and disruption as a result [8].

Sulistyowati et al. [21] proposed a model to reduce the risk of security breaches with the combination of the OCTAVE framework and ISO 27001. The risk assessment was conducted based on the OCTAVE framework, while the information security control and risk mitigation analysis is based on ISO 27001. The sustainability of the proposed improvement method is based on lost expectancy and return on investment. However, this model focuses solely on the security requirements of information assets and not on data security in HEIs.

Table III summarizes the evidence discussed in this section which highlights that most risk assessment work in HEIs based on OCTAVE, OCTAVE Allegro and risk management is based on the ISO framework. Besides that, the scope of those proposed risk assessment models focuses more on the security of assets in HEIs rather than data security. Therefore, our study aims to explore risk assessment based on the NIST SP 800-30 and ISO 27005 framework.

III. PROPOSED MODEL

The proposed model is based on ISO 27005 framework for context establishment and NSIT SP 800-30 framework for risk assessment process. Fig. 1 illustrates the proposed model for this study.

A. Context Establishment

In our study, the context establishment is based on ISO 27005 framework. This process establishes essential criteria for information security management. The context establishment explained the scope and restriction of risk that are adjusted based on the information security level to be achieved.

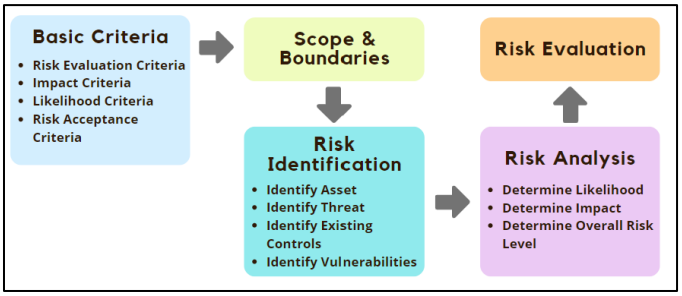


Fig. 1. Proposed Model.

TABLE III. RELATED WORKS ON CYBERSECURITY RISK ASSESSMENT IN HEIS

Authors	Objective	Scope	Framework	Phases
[17]	To assess the Academic Information System asset risk.	Academic Information System asset.	OCTAVE Allegro and ISO 27002	Risk assessment is conducted based on the OCTAVE framework. Implementation of security controls is based on ISO 27002.
[18]	To reduce the risks of a security breach.	Network configuration security	OCTAVE	Phase 1: Identification of weak points in university network configuration. Phase 2: Quantitative risk level measurement for the university's campus network. Phase 3: Enhancement of the university's security position.
[19]	To identify, analyze and manage the risk of academic information systems in HEI using the OCTAVE Allegro method.	Academic Information System	OCTAVE Allegro	Phase 1: Establish drivers Phase 2: Profile assets Phase 3: Identify threats Phase 4: Identify and mitigate risks
[20]	To identify the risk that affects the security of information assets and design some protection strategies for securing those risks.	Assets of Information System	OCTAVE Allegro	Phase 1: Establish drivers Phase 2: Profile assets Phase 3: Identify threats Phase 4: Identify and mitigate risks
[21]	The purpose of the proposed model is to reduce the risk of security breaches. The feasibility of the proposed improvement method is based on lost expectancy and return on investment.	Assets	OCTAVE and IS027001	Phase 1: • Understanding the information security needs. • Identify threats and vulnerabilities. Phase 2: • Identify likelihood. • Identify severity. • Risk assessment. Phase 3: • Analysis of Information security controls based on ISO 27000. • Calculation of loss expectancy. • Remediation plan.

1) Basis Criteria

a) Risk Evaluation Criteria: This study establishes the consideration in evaluating risk with these criteria:

- Confidentiality refers to the safeguarding of data against unauthorized access. NIST defined confidentiality as preserving authorized information access and disclosure restrictions, including safeguards for personal privacy and proprietary information [22]. In this study, when a hacker or other unauthorized individual gains access to a student information system, the students' data has lost its confidentiality.
- Integrity refers to the assurance that the data are unchanged from creation to reception. In this study, loss of integrity occurs when HEI data is accessed or modified by unauthorized parties, resulting in data accuracy and authenticity loss. For example, when a student's data is accessed or modified by a third party, the data's authenticity is not lost.
- Availability means the asset is always available to the authorized user [7]. In this study, loss of availability is defined as the state of an information system being unavailable, resulting in data loss and accuracy. The unavailability could be due to system disruption or malicious attacks by attackers.

b) Impact Criteria: The impact and likelihood of occurrence criteria are determined based on NIST SP 800-30 revision 1, where the rating scale is assessed from 5 being "Very High" to 1 being "Very low" and determined based on CIA triad of Confidentiality, Integrity and Availability. These criteria are presented in Table IV.

c) Likelihood Criteria: The likelihood of occurence criteria and likelihood of threat event resulting in adverse impact are adapted based on NIST SP 800-30 guidelines. Table V shows the likelihood of threat event resulting in adverse impact adapted based on NIST SP 800-30 guidelines. Table VI shows likelihood of threat event resulting in adverse impact.

d) Risk Acceptance Criteria: Risk acceptance is defined as the level of risk taking acceptable to achieve a specific business objective. Determining risk tolerance allows HEI to articulate how much risk the organization is willing to accept [23]. Table VII shows the risk tolerance appetite matrix based on NIST SP 800-30 guidelines.

TABLE IV. IMPACT RATING CRITERIA

Scale	Description	Value
Very High	Unauthorized disclosure of confidential data with a high number of records resulted in an adverse impact on HEIs. The unauthorized modification of the confidential data resulted in data damage or loss which cannot be recovered. The student system is not accessible for more than 24 hours.	5
High	Unauthorized disclosure of confidential data with a medium or low number of records seriously impacted HEIs. The unauthorized modification of the confidential data resulted in data being damaged/ missing, but data can be recovered. The student system is not accessible between 12 hours to 24 hours.	4
Moderate	Unauthorized disclosure of internal data resulted in a moderate impact on HEIs. The unauthorized modification of the internal data resulted in data being damaged/missing but can be recovered. The student system is not accessible between 2-12 hours	3
Low	Unauthorized disclosure of public data resulted in a low impact on HEIs. The unauthorized modification of the internal data resulted in data being damaged/missing but can be recovered. The student system is not accessible between 1-2 hours.	2
Very low	Unauthorized disclosure of unclassified data resulted in a low impact on HEIs. The unauthorized modification of the internal data resulted in data being damaged/missing but can be recovered. The student system is not accessible for less than 1 hour.	1

TABLE V. LIKELIHOOD CRITERIA

Scale	Frequency Number of a Possible Occurrence	Value
Very high	Between 20 to 30 times a year	5
High	Between 10-20 times a year	4
Moderate	Between 5 to 10 times a year	3
Low	Between 2 to 5 times a year	2
Very low	Less than 2 times a year	1

TABLE VI. RESULTING IMPACT SCALE

Scale	Impact Description	Value
Very High	Definitely give a negative impact	5
High	Almost certainly give a negative impact	4
Moderate	A medium probability gives a negative impact	3
Low	A small probability gives a negative impact	2
Very low	Very unlikely to have a negative impact	1

TABLE VII. RISK TOLERANCE MATRIX

Risk level	Impact Description	Scale
Low and Very Low	Risks are acceptable	1 - 4
Medium	Risks can be mitigated	5 - 15
High and Very High	Must be mitigated	15 - 25

2) Scope and boundaries: The scope of the risk assessment determines what will be considered in the assessment and what risk scenarios HEIs could anticipate. Risk assessment scope affects the range of information available to make risk-based decisions and is determined by the organizational official requesting the assessment and the risk management strategy. HEIs risk is not limited to information systems and security but includes financial, strategic, technological, and reputational risks [9]. In this study, our scope covers five types of risks as follows:

a) Strategic Risk: Strategic risk is related to corporate risk. It impacts the development and implementation of an organization's strategy. Strategic risk influence the organization's ability to achieve its long-term goals and objectives [13]. To effectively learn and adapt to new changes, top management needs to carefully define and implement a strategy. When a university implements a new strategy for its business process, the risk associated with that strategy should be considered. Since the COVID 19 pandemic, HEIs have shifted their teaching delivery from physical to online. If staff and students do not adapt to the new environment, the teaching procedures and academic achievement may deteriorate.

b) Operational Risk: The operational risk focuses on managing the risk that occurs in daily operations [9]. It is an occurrence that affects the organizations' ongoing management processes and procedures. Meanwhile, operational risk is defined by Panchal as the likelihood of human error or fraud in manual or automated environments. It also refers to potential threats to an institution's administrative process [11]. Inefficient or defective internal processes, people, control, system, or external events are the causes of business failures. For example, when a new learning management system is implemented in HEIs, teaching and learning activities are modified. If the changes are not effectively implemented, they may severely influence the ongoing student learning process, caused to system downtime and failure.

c) Compliance Risk: Compliance risk is concerned with the adherence to externally imposed laws and regulations, as well as internally bound policies and procedures concerning safety, conflicts of interest, and other issues. [20]. It is associated with conformance to federal, state, and regional rules and regulations [11]. It is concerned not only with externally imposed laws and regulations but also with internal policies and practices. This study investigates compliance risk in relation to research activities undertaken in an academic institution. The institution's research department must follow the laws and regulations of both the university and the government. Failure to comply with or violate applicable laws might result in severe penalties and accreditation revocation.

d) Financial Risk: Financial risk is associated with an initial assessment of HEIs revenues and expenditures and how to manage them [21]. Asset loss, conflict of interest, and

technological risks are financial management or transaction events that harm an organization's profitability and efficiencies. In this study, financial risk refers to the negative consequence of a cyberattack. Attackers can steal sensitive information, disable critical system access, and demand payment before restoring access. They have also threatened institutions with the publication or stolen critical information if they disagree with their requests. Some organizations must pay a ransom to regain access and recover lost data and systems. The sum paid may reduce the university's budget or create insolvency, resulting in insufficient cash for other operations such as research, teaching, maintenance, and development.

e) Reputational Risk: Reputational risks are related to an organization's brand or public image and emerge from the organization's inability to handle any other type of risk accurately [20]. It also includes the external perception of the organization's reputation. Reputational risk is frequently seen as a critical issue [13]. Political difficulties or unconstructive occurrences are examples of events that harm an institution's reputation and public view. The impact of external perception on an institution's image and brand is the focus of reputational risk [11]. This risk may occur due to an institution's failure to manage any or all of the other risks effectively. HEIs must protect their valuable data, assets, and images from sustaining the university's trust among students, parents, alumni, and the general public. Failure to successfully manage this risk will harm the university's reputation, the inability to meet the target of student enrollment, and the failure to meet the target of business and research initiatives.

B. Risk Assessment

In this study, the risk assessment process will be based on NIST SP800-30 because the guidelines contain detailed criteria to analyze the risk.

1) Risk Identification

a) Identify Asset: Typically, a risk assessment encompasses all the organization's critical assets that directly impact the confidentiality, integrity, and availability of the organization's information resources [13]. Table VIII shows the example of information assets in the student information system.

b) Identify Threats: NIST [24] defined a threat as any circumstance or event that has the potential to negatively affect the organization, individuals, other organizations, or the nation's operations and assets via an information system through unauthorized access, destruction, disclosure, or modification of information, and/or denial of service caused by threat sources. In this study, a threat is defined as a potential cause of an adverse event that may harm the HEI environment. Table IX shows an example of threat listings based on NIST SP-800 threat catalogues.

c) Identify Existing Control: The primary aim of this process is to consider both existing and proposed controls when determining the chance that a threat source would exploit the vulnerability. Hence, the more effective the control, the less likely a weakness would be exploited and vice versa.

d) Identify Vulnerabilities: This activity focused on identifying vulnerabilities that the identified threats could exploit. Examples of threat vulnerabilities scenarios are presented in Table X based on NIST SP 800-30 vulnerabilities catalogue.

TABLE VIII. EXAMPLE OF INFORMATION ASSETS IN THE STUDENT INFORMATION SYSTEM

Category	Information Asset
Student information system	Personal sensitive information. Student financial information. Student academic details. Student accommodation details. Study records of course completion and achievements.

TABLE IX. THREAT LISTING

Threat Agent	Threat Action
Students	Possible weak passwords due to lack of password complexity controls
Malicious insiders	System intrusion and unauthorized system access.
Hackers	Send phishing e-mails requesting students to enter their confidential details.

TABLE X. THREAT VULNERABILITIES SCENARIOS

Threat Agent	Threat Action	Vulnerabilities
Students	Open an e-mail requesting sensitive information or click on a malicious link that unknowingly downloads malware onto their device.	Lack of anti-virus and malware prevention.
Malicious insiders	System intrusion and unauthorized system access.	Weak password or due to lack of password complexity.
Hackers	Send phishing e-mails requesting students to enter their confidential details.	Insufficient security awareness and best practices.

2) Risk analysis: Risk analysis is about analyzing the elements that make up each risk scenario to determine [24]:

- The overall likelihood of a risk scenario occurring is calculated based on the combination of the likelihood that the event will occur and the likelihood that the event will have a negative impact.
- The impact (i.e., magnitude of harm) resulting from the occurrence of a risk scenario.

Table XI shows an assessment scale based on the NIST SP800-30 guideline to determine the overall likelihood.

The final risk rating is determined based on the intersection of the impact and overall likelihood for each identified threat and vulnerability pair. The formula to evaluate the risk is:

Risk = Overall Likelihood x Impact.

TABLE XI. OVERALL LIKELIHOOD

Likelihood of threat event initiation occur	Likelihood of threat event results in adverse impact				
	Very Low (1)	Low (2)	Moderate (3)	High (4)	Very High (5)
Very High (5)	Very Low	Moderate	High	Very High	Very High
High (4)	Very Low	Moderate	Moderate	High	Very High
Moderate (3)	Very Low	Low	Moderate	Moderate	High
Low (2)	Very Low	Low	Low	Moderate	Moderate
Very Low (1)	Very Low	Very Low	Very Low	Low	Low

Table XII depicts the risk appetite matrix used to determine risk. If the risk scores are in the black range, the risk is considered high. Meanwhile, if the risk falls into a grey shade, it is classified as moderate risk. If the risk is in the white shade, then the risk is categorized as low risk.

TABLE XII. RISK APPETITE MATRIX

Impact	Overall likelihood				
	1	2	3	4	5
1	1	2	3	4	5
2	2	4	6	8	10
3	3	6	9	12	15
4	4	8	12	16	20
5	5	10	15	20	25

3) Risk evaluation: Lastly, the derived risks will be evaluated according to the risk matrix score and compared to the risk tolerance level specified in the risk criteria. The output will take the next course of action to keep the risks within the organization's risk tolerance level.

IV. DISCUSSION

The core to effective university risk management is cybersecurity risk assessment. It is critical to select a suitable risk assessment approach that may give universities a range of instruments to identify unforeseen events and mitigate the impacts. We conducted extensive literature studies by evaluating existing risk assessment frameworks and related works on risk assessment in HEIs.

Based on our findings, we can conclude that the most dominant risk assessment literature in HEIs utilizes OCTAVE and OCTAVE Allegro framework for risk assessment and ISO 27005 framework for risk management. Hence, our study aims to explore ISO 27005 and NIST SP 800-30 frameworks to establish a risk assessment model for HEIs.

The context establishment and criteria are adapted based on ISO 27005 because it describes how to represent an incident process in risk scenarios. HEIs can assess the likelihood and impact that occurs in the scenarios of information risk to information security with the aid of incident description of risk scenarios. Meanwhile, the risk assessment process is based on the NIST SP 800-30 framework since it includes criteria, scoring, and decision matrices for analyzing risk, whereas ISO 27005 solely focuses on objectives, guidelines and concepts.

V. CONCLUSION AND FUTURE WORKS

This study aimed to establish a cybersecurity risk assessment model for HEIs by modeling the factors associated with HEIs. The method is based on the prominent ISO 27005 and NIST SP 800-30 frameworks. The primary goal of risk assessment in HEIs is to measure the risks and to improve their decision-making in managing the risk within the environment. A proposed cybersecurity risk assessment model was developed, demonstrating that several critical scenarios may arise in the HEIs environment. After evaluating the identified risks, the next step is to identify and determine the next course of action to keep the risks within the organization's risk tolerance level. Future research initiatives could further enhance the proposed model on establishing appropriate countermeasures for risk treatment in the HEI environment.

ACKNOWLEDGMENT

The Universiti Tenaga Nasional funds this research under the BOLD 2021 grant.

REFERENCES

[1] K12 Cyber Secure, "The K-12 Cybersecurity Resource Center The K-12 Cyber Incident Map," The K-12 Cyber Incident Map, 2021. https://k12cybersecure.com/map/ (accessed Mar. 27, 2021).

[2] A. Yeoh, Q. Tariq, and S. Menon, "UiTM students ' data allegedly stolen," The Star, 2019. https://www.thestar.com.my/news/nation/2019/01/26/uitm-students-data-allegedly-stolen-classified-records-compiled-over-18-years-believed-taken-from-va/.

[3] W. Zamora, "Trojans, ransomware dominate 2018–2019 education threat landscape," Malwarebytes Labs, 2019. https://blog.malwarebytes.com/trojans/2019/08/trojans-ransomware-dominate-2018-2019-education-threat-landscape/ (accessed Mar. 27, 2021).

[4] Kaspersky Team, "Student surprise : Malware masked as textbooks and essays Download an essay , get some malware thrown in Which types of malware are disguised as textbooks and essays ?," Karpersky Daily, 2019. https://www.kaspersky.com/blog/back-to-school-malware-2019/28316 / (accessed Mar. 27, 2021).

[5] S. Williams, "Cyber criminals target education sector as remote learning increases," Security Brief, 2021. https://securitybrief.eu/story/cyber-criminals-target-education-sector-as-remote-learning-increases (accessed Mar. 27, 2021).

[6] R. Sani, "Curbing cyber threats in online learning," New Straits Times Times, 2020. https://www.nst.com.my/education/2020/05/592083/curbing-cyber-threats-online-learning.

[7] A. S. Sendi, M. Jabbarifar, M. Shajari, and M. Dagenais, "FEMRA: Fuzzy expert model for risk assessment," 2010, doi: 10.1109/ICIMP.2010.15.

[8] P. Panchal, "Information Technology Risks in Higher Education: Strategy for Assessment, Planning and Management," CIO Review, 2022. https://education.cioreview.com/cioviewpoint/information-technology-risks-in-higher-education-strategy-for-assessment-planning-and-management-nid-4585-cid-27.html.

[9] S. S. Hassen and M. S. Zakaria, "Managing university IT risks in structured and organized environment," Res. J. Appl. Sci. Eng. Technol., vol. 6, no. 12, pp. 2270–2276, 2013, doi: 10.19026/rjaset.6.3858.

[10] D. Rios Insua, A. Couce-Vieira, J. A. Rubio, W. Pieters, K. Labunets, and D. G. Rasines, "An Adversarial Risk Analysis Framework for Cybersecurity," Risk Anal., vol. 41, no. 1, pp. 16–36, 2019, doi: 10.1111/risa.13331.

[11] C. J. Alberts, S. G. Behrens, R. D. Pethia, and W. R. Wilson, "Operationally Critical Threat, Asset, and Vulnerability Evaluations (OCTAVE(SM)) Framework, Version 1.0. Carnegie Mellon Software Engineering Institute," 1999. [Online]. Available: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=13473.

[12] R. A. Caralli, J. F. Stevens, L. R. Young, and W. R. Wilson, "Introducing OCTAVE Allegro: Improving the Information Security Risk Assessment Process," 2007. [Online]. Available: https://resources.sei.cmu.edu/asset_files/technicalreport/2007_005_001_14885.pdf.

[13] M. Talabis and J. Martin, Information Security Risk Assessment Toolkit: Practical Assessments through Data Collection and Data Analysis. Syngress, 2013.

[14] T. R. Peltier, Facilitated Risk Analysis Process (FRAP). Auerbach Publications, 2001.

[15] A. Refsdal, B. Solhaug, and K. Stølen, "Cyber-Risk Management," in Cyber-Risk Management. SpringerBriefs in Computer Science., Springer, 2015.

[16] N. A. Hashim, Z. Z. Abidin, N. A. Zakaria, R. Ahmad, and A. P. Puvanasvaran, "Risk assessment method for insider threats in cyber security: A review," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 11, pp. 126–130, 2018, doi: 10.14569/ijacsa.2018.091119.

[17] M. T. Jufri, M. Hendayun, and T. Suharto, "Risk-assessment based academic information System security policy using octave Allegro and ISO 27002," Proc. 2nd Int. Conf. Informatics Comput. ICIC 2017, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/IAC.2017.8280541.

[18] C. Joshi and U. K. Singh, "Information security risks management framework – A step towards mitigating security risks in university network," J. Inf. Secur. Appl., vol. 35, pp. 128–137, 2017, doi: 10.1016/j.jisa.2017.06.006.

[19] J. Hom, B. Anong, K. B. Rii, L. K. Choi, and K. Zelina, "The Octave Allegro Method in Risk Management Assessment of Educational Institutions," Aptisi Trans. Technopreneursh., vol. 2, no. 2, pp. 167–179, 2020, doi: 10.34306/att.v2i2.103.

[20] J. S. Suroso and M. A. Fakhrozi, "Assessment of Information System Risk Management with Octave Allegro at Education Institution," Procedia Comput. Sci., vol. 135, pp. 202–213, 2018, doi: 10.1016/j.procs.2018.08.167.

[21] I. Sulistyowati and R. V. H. Ginardi, "Information Security Risk Management with Octave Method and ISO/EIC 27001: 2013 (Case Study: Airlangga University)," IPTEK J. Proc. Ser., vol. 0, no. 1, pp. 32–38, 2019.

[22] NIST, "Guide for Conducting Risk Assessments," 2012. [Online]. Available: http://dx.doi.org/10.6028/NIST.SP.800-30r1.

[23] ISACA, "ISACA, CRISC Review Manual 6th Edition," 2015. CSA, "Guide to Conducting Cybersecurity Risk Assessment for Critical Information Infrastructure," no. December, 2019.

Acne Classification with Gaussian Mixture Model based on Texture Features

Alfa Nadhya Maimanah, Wahyono*, Faizal Makhrus
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia

***Abstract*—This paper presents an acne detection method on face images using a Gaussian Mixture Model (GMM). First, the skin area in the face image is segmented based on color information using the GMM. Second, the candidates of the acne region are then extracted using a Laplacian of Gaussian-based blob detection strategy. Then, texture features are extracted from acne candidates using either a Gabor Filter or Gray Level Co-occurrence Matrix (GLCM). Lastly, these features are then utilized as input in the GMM for verifying whether these regions are acne or not. In our experiment, the proposed method was evaluated using face images from ACNE04 dataset. Based on the experiment, it is found that the best classification results were obtained when GLCM features in the Cr-YCbCr channel are applied. In addition, the proposed method has competitive performance compared to K-Nearest Neighbor (KNN).**

Keywords*—Acne; GLCM; Gabor filter; Gaussian mixture model

I. Introduction

In recent years, the beauty technology industry has experienced significant growth along with increasing public enthusiasm and awareness of skin health. One major concern of skin problems is *acne vulgaris*. Acne occurs when sebum is trapped in the hair follicles and attacked by *Propionibacterium acnes* [1]. Adolescents to adults may experience this problem. It could leave scars that influence the level of confidence of some people. A common method to determine the acne severity is through manual counting by a dermatologist. The process is susceptible to subjectivity factors, both inter-observers—the same patient has different assessments by different dermatologists—and intra-observer—the same patient has different assessments by the same dermatologist on different days [1]. In addition, the technique is not effective in terms of time and effort spent by a dermatologist [2]. Therefore, technology is needed to assist the process of acne severity assessment through face images.

There have been several developed technologies which are employed to assess facial skin. In addition to helping determine the appropriate and accurate type of treatment, the presence of technology can also attract the attention of customers. The types of technology developed are quite diverse, ranging from instruments equipped with skin analysis systems such as VISIA, Internet of Things (IoT) such as LG U+ LTE Magic Mirror, mobile-based applications such as TroveSkin, and Skin Genius by L'Oréal Paris, as well as Software as a Service (SaaS) such as Haut.AI. Facial skin analysis models are developed based on Artificial Intelligence (AI) which is trained using the data, the company had collected.

This research proposes a study of comparison of texture-based feature extraction methods for assessing the acne severity on human face images. To the best of our knowledge, the research related to acne images generally aims to segment and/or classify acne types. However, research that aims to determine the severity of acne suffered by patients is still limited. Research [3] contributed to determining the severity of acne, but it was still based on the area of acne on the right and left cheeks without using appropriate standards. Since there are still few studies on acne images with standardized acne severity, this research used the criteria formulated by Hayashi [4], which was also utilized in the dataset developed by Wu et al [5]. This criterion estimates the acne severity based on the number of papules and pustules detected on a face image captured with an angle of 70° from the front side.

Furthermore, this research utilizes texture-based features, namely GLCM and Gabor Filter. GLCM was chosen because it was inspired by research related to the acne types classification conducted by Ramadhani [6] which achieved an accuracy value of 72%. In addition, research conducted by Chang and Lio [7] successfully detected acne on face images using GLCM features with accuracy of 99.40%. It was proved that GLCM has the ability to extract the features needed for acne detection. On the other hand, Gabor Filter was chosen because research conducted by Jeon and Cheoi [8] could successfully detect abnormal areas on skin images, including small acne and regions with low contrast levels, which open the possibility of implementing this method for acne detection. To classify acne and non-acne areas, this research conducted an experiment to use the Gaussian Mixture Model (GMM) method. This method is usually used as a density estimator so it is suitable for clustering problems. In addition, the GMM has also been proved as a good approach for classification which was implemented by Dey [9] for skin classification as well as Wan et al. [10] for classifying 10 kinds of datasets from the UCI Machine Learning Repository.

The rest of this paper is organized as follows. Section II discusses the research related to this topic. Section III explains the details of the proposed method. Section IV demonstrates the experimental result. Conclusions are given in Section V.

II. Literature Review

Based on the feature extraction method, the research related to acne images is generally divided into two categories: hand-crafted and deep learning-based feature extractions. The features from the hand-crafted strategies can be obtained based on color, shape, and texture [11]. A comprehensive experiment and observation are required to

*Corresponding Author.

determine the best features as the basis for acne detection, unlike deep learning where the model can extract its own features automatically.

Deep learning is the latest research trend regarding images with acne objects. Zhao et al. [12] used a regression model with transfer learning on ResNet-152 to determine the acne severity but it did not achieve a good performance since the data were imbalanced, which is dominated by one class. Arifianto and Muhimmah [13] used transfer learning on ResNet-50 to detect acne and obtained an accuracy of 63.2%. Both studies faced the problem of limited data with good quality which resulted in low accuracy. Junayed et al. [14] also utilized deep learning for classifying five classes of acne with accuracy over 94%, but it required an expensive computation time.

On the other hand, the hand-crafted feature extraction based on either color or texture was also implemented. Acne color features tend to be inconsistent with variations in lighting and skin color on the same type of acne as in [15] and [16]. With the input of the cropped acne area from the face image, the author [6] utilizes feature textures for classification of acne types. Authors in [7] extracted Gray Level Co-occurrence Matrix (GLCM) statistical features on a whole face image which was then divided into blocks. However, the image acquisition process required a special device with standard camera settings, lighting, and shooting distance parameters that had been determined beforehand. It was less flexible to be implemented as a mobile application only has a smartphone camera available. Nevertheless, the hand-crafted approach could achieve a high accuracy of 99.40% with less computation time compared to deep learning approach. Therefore, this research would be conducted to study more on the utilization of hand-crafted texture-based features to detect acne on face images.

The input images used in previous studies related to acne recognition were quite diverse. Some studies used whole human face images with various levels of acne severity, as in the research [12] and [13], while other studies used cropped images on the acne object only, such as [14] which used a dataset from Dermnet consisting of five types of acne.

Furthermore, research related to acne images can be grouped into three objectives: acne segmentation, acne classification, and acne severity assessment. Research conducted by Maroni et al. [2] detected and counted the number of acnes with a sequence of processes starting from body part detection, skin segmentation, heatmap creation, acne extraction, and blob detection. The best skin segmentation was generated by the Random Forest method based on the 15 most informative features explored. Acne extraction was carried out using the adaptive thresholding of heatmap images. After that, acne was detected and counted using the Laplacian of Gaussian (LoG) filter. Research conducted by Jeon and Cheoi [8] aimed to cluster abnormal areas on human skin using the density-based spatial clustering of applications with noise (DBSCAN) algorithm based on the features obtained from the Gabor Filter. In their study, the skin segmentation was not carried out first even though there was an image input in the form of a face. The proposed method had better performance than the [17] method because it could detect small acne and regions with low contrast levels.

Lastly, some researchers classified acne based on texture features. Ramadhani [6] utilized GLCM to obtain texture characteristics including contrast, energy, entropy, correlations, and dissimilarity. Among these characteristics, entropy was the most influential statistical feature since it represents texture irregularities. The overall accuracy value obtained was 72%. Some studies classified acne based on color features, like a research by Darmawan et al. [15] where RGB color intensity was used. Although it had been able to detect types of acne, there were still limitations due to lighting factors during image acquisition resulting in color values discrepancy and accuracy. Gunawan et al. [16] conducted segmentation using Region Growing and classified the acne types using Self Organizing Map (SOM). RGB histogram feature was used as SOM input. The classification accuracy was still not ideal due to acne color variations influenced by diversity of skin color and lighting conditions in each image. Several studies used the segmentation results to classify the types of acne, like in Arora and Sarvani [17] who explored the methods of acne segmentation and machine learning models for acne classification. Compared to color and texture segmentation, the 2-level K-means clustering had the best accuracy at 70%. While the classification of acne and acne scars had an average accuracy of 80% using the Fuzzy C-Means (FCM).

III. METHODS

A. Classification Model Development

A total of 40 images from the ACNE04 dataset—10 images from each Hayashi Criteria—are used for evaluation. These images have non-uniform dimensions therefore they are resized to 320×320. The bounding box coordinates in the Extensible Markup Language (XML) annotation document are also adjusted to the same dimension which is then used for image cropping to obtain acne blocks. On the other hand, for non-acne blocks, a whole image is divided into a uniform block size of 20×20 which is then curated manually to remove the ones with acne. In total, there are 908 acne blocks and 870 non-acne blocks.

The first texture features are extracted from GLCM. The distance and angles chosen are one pixel and (0°, 45°, 90°, and 135°), respectively. For comparison purposes, this feature extraction is carried out on four different channels of color space, i.e. Grayscale, Hue-HSV, Red-RGB, and Cr-YCrCb. There are six features calculated from the GLCM including contrast, dissimilarity, correlation, energy, homogeneity, and ASM, each of which was the average of the four neighboring angles. These features are then normalized to the range of [0,1]. Finally, there will be six GLCM features obtained from each block. The second texture features are extracted from Gabor filtered images. The two-dimensional Gabor filter is a Gaussian kernel function modulated by a sinusoidal wave [18]. The Gabor filter bank is created with the size of 3×3 by adjusting the five variables in formula 1. The first three variables followed the research by [8]: $\Psi=\pi/2$; $\lambda=0.8$; $\gamma=0$. For comparison purposes, the last two variables are set on two different configurations: $\theta=0°$, 30°, 45°, 60°, 90°, 120°, 135°,

150°; σ=1 and θ= 0°, 45°, 90°, 135°; σ=1,2. The visualization of the Gabor filters used is displayed in Fig. 1. After the filtering process, two statistical features, namely, mean and variance are calculated. Finally, there will be 16 Gabor features obtained from each block. After both features are gathered, they are trained for acne and non-acne classification using GMM.

$$g(x, y; \Psi, \lambda, \gamma, \theta, \sigma) = exp\left(\frac{-x'^2+\gamma^2 y'^2}{2\sigma^2}\right) cos\left(2\pi \frac{x'}{\lambda} + \psi\right) \quad (1)$$

The GMM classifier performs acne and non-acne classification using two GMMs, one model for acne features and the other for non-acne features. After training those models, a new testing feature is classified as acne if the log-likelihood value in acne GMM is higher than in non-acne GMM. To determine the optimal number of Gaussian components in the two GMMs, experiments are carried out on the number of Gaussian components in the range of 2-20 so that the lowest Bayesian Information Criterion (BIC) value could be obtained. Below is the pseudocode to create a GMM Classifier.

Input: GLCM/Gabor Features, K_{max} = 20, and C = 2 (acne and non-acne)
Output: K_{opt} for every class c
for c=0:C do
  for k=2: K_{max} do
    Apply GMM−EM with k number of Gaussian components;
    Calculate BIC based on maximum parameters obtained;
  end for
  K_{opt} = arg min k (BIC)
end for

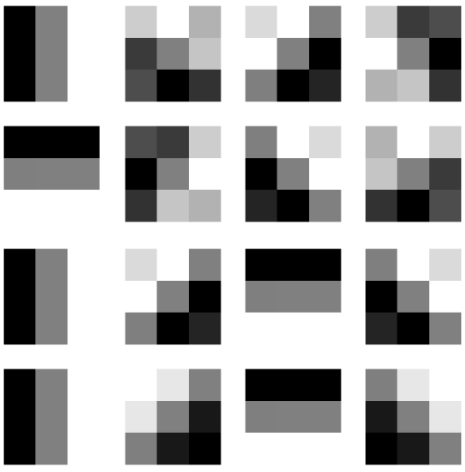


Fig. 1. Gabor Filter Bank: 8 Angles-1 Standard Deviation (Row 1 and 2), 4 Angles-2 Standard Deviations (Row 3 and 4)

B. Skin Segmentation

To evaluate the classification model, skin segmentation steps are implemented as shown in Fig. 2. A total of 20 images from the ACNE04 dataset—five images from each Hayashi Criteria—are used for testing. These images are also resized to 320×320 dimensions. Then an enhancement process is carried out on the a* CIELab channel. The reason behind this is that the a* channel represents the level of pixel redness that is independent of lighting, making it stronger for detecting acne that tends to be redder than the surrounding skin. The enhancement started with the unsharp mask to sharpen the image after going through the resizing step. Then, the Difference of Gaussian (DoG) process is applied by subtracting the unsharp masked image from the Gaussian Blurred image following the research conducted by [19]. The resulting image highlights the pixels that tended to be red. The Gaussian Blur filter parameters use kernel size of 19×19 and σ=13. The next step is skin segmentation. For comparison purposes, two skin segmentation methods are used. The first is GMM segmentation based on BGR skin and nonskin pixel values from the Skin Segmentation Dataset, UCI Machine Learning Repository. The second method is Otsu Thresholding on the median blurred Cr image. This channel is selected since it is a red chromatic channel that could help the process of blurring pixels whose intensity did not resemble the skin—in which the color tends to be reddish. The size of the median blur filter is 21×21. The result of Otsu Thresholding is then used as the mask on the resulting image from a* CIELab enhancement step. Face parts such as the mouth, right eye, left eye, right eyebrow, and left eyebrow could be susceptible to being misdetected as acne, especially those with similar features in terms of color intensity, such as the mouth. Therefore masking is done on these parts. Face parts detection was done with the help of the DLIB library. The resulting image is then being used as a mask. The next step is to do acne candidate thresholding on the already masked a* CIELab enhanced image with the threshold value of 128. To remove the noise, a morphological opening operation is performed with a kernel size of 3×3.

C. Acne Detection

To calculate the number of acne candidates, the blob detection method is used with the Laplacian of Gaussian (LoG) kernel. The parameters used include the minimum standard deviation of the Gaussian kernel = 1, the maximum standard deviation of the Gaussian kernel = 5, the number of intermediate values, standard deviation = 15, threshold = 0.2, and overlap = 0.1. Each detected blob stored the coordinates of the blob's center (x,y) and its radius. They are used to determine the bounding box coordinates of the acne candidates which are then used for the cropping process. The cropping process is carried out by taking the pixels in the bounding box's coordinates range. The cropped pixel blocks are stored for texture features calculation. The features obtained from all blocks of acne candidates are used to predict the acne classification using the GMM Classifier model that had been trained previously. To reduce the number of overlapping bounding boxes while recognizing the same object, the Non-Maximum Suppression (NMS) process is carried out. The output image is the original image with a bounding box on each of the detected acne. In addition, the text of the number of detected acne and the severity are also displayed.

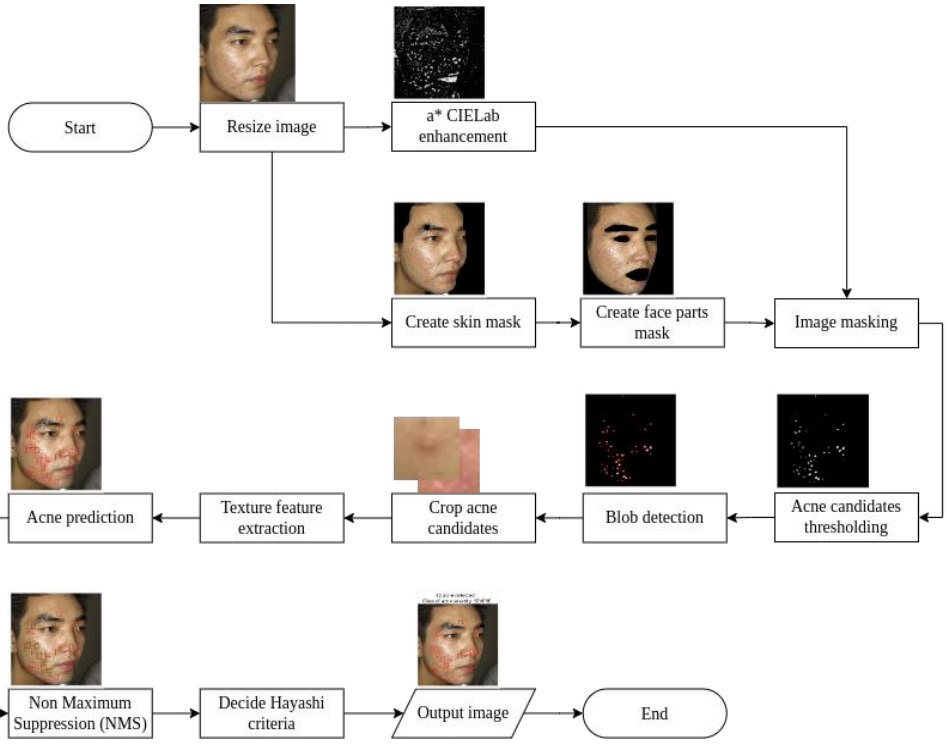


Fig. 2. Flowchart of the Proposed System.

IV. Experimental Results and Discussions

A. Skin Segmentation Result

Skin segmentation using the GMM method can separate facial skin areas in some images although hair, background, and clothing areas are still visible. Based on observations, these non-facial areas occur because of the morphological operation process carried out to remove noise and holes from the initial segmentation results. Several kernel sizes (3×3,5×5,7×7), the number of iterations (1,2,5,7), and the order of operations morphology are tried, however the obtained results still include 20-30% of non-facial regions. Fig. 3 on the left is the result of skin segmentation using the GMM method by performing closing and dilation operations using a kernel size of 5×5 and iterating 7 and 2 times respectively. Fig. 3 on the right is the result of skin segmentation using Otsu Thresholding on Median Blurred Cr Image which is better than the GMM method based on the minimum visible hair area, almost no visible background area, and minimum visible clothing area factors. The computation time of this second method is faster with an average computation time of 0.005 seconds, compared to the GMM method with an average computation time of 0.419 seconds. Therefore, based on the quality of the segmentation results and computational time, the experiments use Otsu Thresholding on Median Blurred Cr Image.

B. Classification Result with GLCM Features

Acne classification using the GMM method with the input of GLCM features is shown in Table I. In addition, Table II presents the details of Table I. Viewed from the prediction accuracy, all channels have accuracies below 55%. The highest total performance is obtained by Red-RGB channel based on the validation and test accuracies and the second place is Grayscale channel. Meanwhile, viewed from the precision value, all channels have relatively low precision in the range of 0.45-0.52, which means that there are still quite a lot of false positives. On the other hand, based on the recall value, Cr-YCrCb channel has the highest value at 0.78 that it becomes the best among the other three channels in correctly predicting acne close to the ground truth. Cr-YCrCb also has the best F1-Score value of 0.56.

Fig. 3. Example of Skin Segmentation Result with GMM Segmentation (Left) and Otsu Thresholding on Median Blurred Cr Image (Right)

TABLE I. Comparison of Testing and Prediction Accuracy with GLCM Features

Channel	K of non-acne	K of acne	Val Acc	Test Acc
Grayscale	10	7	95.22%	48.89%
Red-RGB	6	5	94.38%	53.20%
Cr-YCrCb	11	12	87.92%	45.14%
Hue-HSV	11	15	71.34%	41.75%

TABLE II. MODEL EVALUATION WITH GLCM FEATURES

Channel	Acc	Prec	Rec	F1	Time	IoU	Hayashi Correct
Grayscale	48.89%	0.52	0.58	0.50	1.228	0.466	10
Red-RGB	53.20%	0.53	0.53	0.49	1.209	0.442	10
Cr-YCrCb	45.14%	0.50	0.78	0.56	1.260	0.463	11
Hue-HSV	41.75%	0.45	0.70	0.52	1.236	0.452	12

Based on the number of images where the Hayashi class [4] is determined correctly, Hue-HSV is the best with 12 images. Having the recall value of 0.70, Hue-HSV channel performs below the Cr-YCrCb because more acne was still predicted as non-acne (it has more false negatives). In terms of computation time, the four channels do not differ much in the range of 1.2 seconds since the computational load tends to be the same as seen from the number of calculated features and steps. Based on the Intersection over Union (IoU) values [20], all four channels are above 0.4 which mean that they are quite good at detecting the location of acne. Nonetheless, since the size of the ground truth bounding boxes varies while the size of the prediction bounding boxes is uniform, the IoU value is difficult to approach 1. From all the experiment results, Cr-YCrCb is chosen as the best performance of GMM classification since it has the highest recall value of 0.78 although its accuracy value is low—in third place. An example of the output images can be seen in Fig. 4. The recall metric is chosen since in this problem, it is more important to correctly identify positive acne (the fewer false negatives are the better). Cr-YCrCb still has many errors in detecting non-acne, as indicated by the high number of false positives (low accuracy). Although in this case, this error is not life-threatening, it is much better to reduce it.

C. Classification Result with Gabor Features

Table III shows the comparison of validation accuracy and testing accuracy of GMM classification using Gabor features. The detailed evaluation results are also shown in Table IV. Viewed from the prediction accuracy, all filters do not achieve high accuracy. The highest was obtained by 4 degrees-2 standard deviations (4deg2sd) filter variation at 55.43%. Meanwhile, viewed from the precision value, both still have low values around 0.5 which mean that there are still many false positives. Compared to the recall these two filters do not differ too much around 0.5. Therefore, they are not good at predicting the correct acne. In addition, since both precision and recall from those two filters have small differences then the F1-Score values differ by only 0.02.

These two filters have the same number of images where the Hayashi class is determined correctly. It means that even though the number of detected acne is close to ground truth, the false negative predictions are still a lot. In terms of the computational time, they are both in the range of 0.4 seconds. The reason for the computational load which tends to be similar is because of the similarity in the number of features and steps. Based on the IoU value, the results show that both are above 0.4 which means that they are quite good at detecting the location of acne. Nevertheless, since the size of the ground truth bounding boxes varies while the size of the prediction bounding boxes is uniform; hence the IoU value is difficult to approach 1. Based on these two experiments, the 8 degrees-1 standard deviation (8deg1sd) filter parameter is chosen as the best filter since it has the highest recall value of 0.54.

TABLE III. COMPARISON OF TESTING AND PREDICTION ACCURACY WITH GABOR FEATURES

Filter	K of non-acne	K of acne	Val Acc	Test Acc
4deg2sd	3	3	67.69%	55.43%
8deg1sd	3	3	67.82%	54.27%

TABLE IV. MODEL EVALUATION WITH GABOR FEATURES

Filter	Acc	Prec	Rec	F1	Time	IoU	Hayashi Correct
4deg2sd	55.43%	0.52	0.51	0.48	0.449	0.446	10
8deg1sd	54.27%	0.54	0.54	0.50	0.472	0.411	10

Fig. 4. Example of the Output Images from the Proposed Method (Detected Acnes Represented by Red Bounding Boxes).

D. GLCM Features vs Gabor Filter Features

An important factor that affects the performance of the model is the quality and ability of the features to represent acne and non-acne characteristics. Viewed from the number of features, Gabor Filter [8] has 16 features while the GLCM has 6 features (initially there are 24, but then only the average value of all neighboring directions of each feature is used [7]). Having fewer features, the GLCM has a generally better recall value than the Gabor Filter although they tend to have lower accuracy. Therefore it is concluded that the GLCM features tend to be better at representing acne and non-acne objects even though in terms of computation time it is longer than the Gabor features. Models with GLCM features require computation time around 1.2 seconds while Gabor features only 0.4 seconds. Another factor that affects the model performance is the classification class which is limited to only two classes, namely acne and non-acne. Based on observations, the model detected several blocks as positive with characteristics close to acne but they are not, such as acne scars, spots, moles, and image lighting that make the skin tend to look red. This causes the high number of false positives in the prediction.

E. Color vs Texture

To determine the performance of the texture features in acne detection, some experiments are conducted. One reason to conduct this research is that color features are not good enough to detect acne since it tends to be inconsistent with variations in lighting and skin color during the image acquisition process even for the same type of acne as stated in [15] and [16]. Therefore the use of the texture features is proposed in order to improve the detection results. However, from the results presented in Table V, it shows out that the color features are still better than the texture features. The color features have the highest recall, F1-Score, and IoU values. With a faster computation time due to less computational load, the accuracy values of color features are generally near to the accuracy values of texture features. Therefore, it is concluded that the use of texture features for acne detection is not better than color features.

TABLE V. TEXTURE FEATURES COMPARED WITH COLOR FEATURES

Feature	Acc	Prec	Rec	F1	Time	IoU	Hayashi Correct
Color	44.15%	0.52	0.79	0.58	0.383	0.473	12
GLCM Grayscale	48.89%	0.52	0.58	0.50	1.228	0.466	10
GLCM Red-RGB	53.20%	0.53	0.53	0.49	1.209	0.442	10
GLCM Cr-YCrCb	45.14%	0.50	0.78	0.56	1.260	0.463	11
GLCM Hue-HSV	41.75%	0.45	0.70	0.52	1.236	0.452	12
Gabor 4deg2sd	55.43%	0.52	0.51	0.48	0.449	0.446	10
Gabor 8deg1sd	54.27%	0.54	0.54	0.50	0.472	0.411	10

F. GMM vs KNN (K-Nearest Neighbor)

To determine the performance of GMM as a classification model, a comparison is made with KNN as one of the supervised machine learning models. The inputs of KNN are GLCM features since they produce better performance than the Gabor Filter features in GMM classification. Based on the number of images where the Hayashi class is correctly determined by using KNN classifier, Hue-HSV is the best channel with 14 correct images. KNN with the Hue-HSV channel is good at predicting acne blocks that are quite close to the ground truth given the recall value of 0.69. However, Cr-YCrCb achieved the highest recall value at 0.77. Despite that, both of those channels have equal F1-Score at 0.54. The computational time of four channels is approximately 1.4 seconds for the reason that the computational loads are similar. Based on the IoU values which are approximately 0.4, they are quite good at detecting the location of acne.

Viewed from the overall accuracy values, the KNN model is better than the GMM. However, the overall recall and F1-Score values are still below GMM except using Cr-YCrCb channel which has a high recall at 0.77. The overall computation time and IoU between KNN and GMM do not differ much. GMM Classifier is better in general at predicting Hayashi class correctly. It can be concluded that GMM as a classification model has competitive performance compared to KNN based on the evaluation parameters. The performances of GMM and KNN with the GLCM features can be seen in Table VI.

TABLE VI. GMM CLASSIFIER AND KNN CLASSIFIER EVALUATION WITH GLCM FEATURES

Channel	Acc	Prec	Rec	F1	Time	IoU	Hayashi Correct
GMM Grayscale	48.89%	0.52	0.58	0.50	1.228	0.466	10
GMM Red-RGB	53.20%	0.53	0.53	0.49	1.209	0.442	10
GMM Cr-YCrCb	45.14%	0.50	0.78	0.56	1.260	0.463	11
GMM Hue-HSV	41.75%	0.45	0.70	0.52	1.236	0.452	12
KNN Grayscale	60.12%	0.57	0.39	0.43	1.322	0.358	7
KNN Red-RGB	59.15%	0.58	0.32	0.39	1.434	0.374	7
KNN Cr-YCrCb	41.76%	0.47	0.77	0.54	1.484	0.473	12
KNN Hue-HSV	49.90%	0.52	0.69	0.54	1.422	0.444	14

All experiments with GMM Classifier have low accuracies in the range of 40-50% since there are still many false positives—or in the other word there are still many parts of the skin that are misdetected as acne. Therefore, the number of detected acne is often bigger than the ground truth. The reason behind this is the limited annotation of the dataset—only acne is labeled—causing the model to fail in recognizing non-acne objects such as acne scars, spots, and moles to acne objects. Adding professional annotations by dermatologists for those

objects may improve the classification performance to let the model learn better.

V. CONCLUSION

The best acne classification result based on recall value is achieved by using the GMM classifier with the GLCM features in Cr-YCrCb channel as the input. The recall is 0.78. It also has a faster computation time, which is about 0.3 seconds compared to the worst method using texture features which is about 1.2 seconds. This classification method is also compared to a standard classification method which is KNN and shows that it outperforms all the evaluation criteria (see Table VI). Some suggestions for further research include the using of a face frame during image acquisition to keep the distance and the captured face size to be uniform, trying other classification algorithms to improve the performance, and increasing the number of training images as well as complementing them with a wider variety of colors and skins.

REFERENCES

[1] R. Ramli, A. Malik, A. Hani, and A. Jamil, "Acne analysis, grading and computational assessment methods: an overview", Skin Research and Technology, vol. 18, no. 1, pp. 1-14, 2012.

[2] G. Maroni, M. Ermidoro, F. Previdi, and G. Bigini, "Automated detection, extraction and counting of acne lesions for automatic evaluation and tracking of acne severity", IEEE Symposium Series on Computational Intelligence 2017, pp. 1-6, 2018.

[3] C. Hsia, T. Lin, J. Lin, H. Prasetyo, S. Chen, and H. Tseng, "System for recommending facial skincare products", Sensors and Materials, vol. 32, no. 10, pp. 3235-3242, 2020.

[4] N. Hayashi, H. Akamatsu, and M. Kawashima, "Establishment of grading criteria for acne severity", The Journal of Dermatology, vol. 35, pp. 255-260, 2008.

[5] X. Wu, W. Ni, L. Jie, Y.K. Lai, S. Cheng, Ming-Ming, and J. Yang, "Joint acne image grading and counting via label distribution learning", IEEE International Conference on Computer Vision, 2019.

[6] M. Ramadhani, "Classification of acne types based on texture using the GLCM method", "Klasifikasi jenis jerawat berdasarkan tekstur dengan menggunakan metode GLCM", e-Proceeding of Engineering, vol. 5, no. 1, pp. 870-876, 2018.

[7] C. Chang and H. Liao, "Automatic facial spots and acnes detection system", Journal of Cosmetics, Dermatological Sciences and Applications, vol. 3, pp. 28-35, 2013.

[8] M. Jeon and K. Cheoi, "Detection of abnormal region of skin using gabor filter and density-based spatial clustering of applications with noise", Journal of Korea Multimedia Society, vol. 21, no. 2, pp. 117-129, 2018.

[9] S. Dey, Python Image Processing Cookbook. Birmingham: Packt Publishing, 2020.

[10] Hu. Wan, H. Wang, B. Scotney, and J. Liu, "A novel Gaussian mixture model for classification", IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 3298-3303, 2019.

[11] M. Dhanashree, S. Kalel, M. Pooja, M. Pisal, M. Ramdas, P. Bagawade, and B. Scholar, "Color, Shape and Texture feature extraction for Content Based Image Retrieval System: A Study", International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 4.

[12] T. Zhao, H. Zhang, and J. Spoelstra. "A computer vision application for assessing facial acne severity from selfie images", ArXiv, vol. 1907.07901, 2019.

[13] J. Arifianto and I. Muhimmah, "Face acne detection web application using deep learning algorithm with TensorFlow", "Aplikasi web pendeteksi jerawat pada wajah menggunakan algoritma deep learning dengan TensorFlow", Prosiding Automata UII, vol. 2, no. 2, 2021.

[14] M. Junayed, A. Jeny, A. Atik, N. Neehal, A. Karim, S. Azam, and B. Shanmugam, "AcneNet - A deep CNN based classification approach for acne classes", 12th International Conference on Information Communication Technology and System, pp. 203-208, 2019.

[15] A. Darmawan, A. Rositasari, and I. Muhimmah, "The identification system of acne type on Indonesian people's face image", IOP Conference Series: Materials Science and Engineering, vol. 803, no. 1, 2020.

[16] A. Gunawan, R. Adipranata, and G. Budhi, "Making acne segmentation and classification applications using region growing and self-organizing map methods", "Pembuatan aplikasi segmentasi dan klasifikasi jerawat dengan metode region growing dan self organizing map", Jurnal Infra, vol. 5, no. 1, pp. 1-6, 2017.

[17] N. Alamdari, K. Tavakolian, M. Alhashim, and R. Fazel-Rezai, "Detection and classification of acne lesions in acne patients: A mobile application", IEEE International Conference on Electro Information Technology, pp. 739-743, 2016.

[18] N. Arora and G. Sarvani, "A review paper on Gabor filter algorithm & its applications", International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), vol. 6, no. 9, 2017.

[19] F. Vasefi, W. Kemp, N. MacKinnon, M. Amini, M. Valdebran, K. Huang, and H. Zhang, "Automated facial acne assessment from smartphone images", Proceeding SPIE Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XVI, vol. 10497, 2018.

[20] R. Padilla, S. Netto, and E. Silva, "A Survey on Performance Metrics for Object-Detection Algorithms", Proceedings of International Conference on Systems Signals and Image Processing, pp. 237-242, 2020.

Learning Content Classification and Mapping Content to Synonymous Learners based on 2022 Augmented Verb List of Marzano and Kendall Taxonomy

S. Celine[1]*
Research Scholar
Department of Computer Science
Sacred Heart College, Tirupattur, India

F. Sagayaraj Fransis[3]
Professor, Department of Computer Science
Pondicherry Engineering College
Pondicherry, India

M. Maria Dominic[2]
Assistant Professor
Department of Computer Science
Sacred Heart College, Tirupattur, India

M. Savitha Devi[4]
Assistant Professor, Department of Computer Science
Periyar University Constituent College of Arts and Science
Harur, India

***Abstract*—Finding suitable learning content for learners with different learning styles is a challenging task in the learning process. Hence it is essential to follow some learning taxonomies to deliver learner-centric learner content. Learning taxonomies are used to express various learning practices and learning habits to be followed by the learner for a better learning process. The investigator has already classified the learners based on the 2022 augmented verb list of Marzano and Kendall taxonomy. The main objective of this paper is to minutely classify the tutor-defined learning contents according to the domains as well as the subdomains of the considered taxonomy which is in text format. Providing personalized learning content could help the learners for a better understanding of learning content and their interrelationship which in turn produce better learning outcomes. Mapping the six levels of learning contents into the corresponding learner is a challenging task. Hence the investigator has chosen seven algorithms including Bagging, XG Boost, Support Vector Machine from Machine Learning and four algorithms including Convolutional Neural Network, and Deep Neural Network in Deep Learning algorithm to classify the learning contents. The experimental results indicate that Support Vector Machine performed well in machine learning and Deep Neural Network yields good performance in deep learning in the learning content classification process. These micro contents were organized using a property graph. Further, the micro contents were retrieved from the property graph using SPARQL for mapping the classified contents to the corresponding learners to achieve personalization in the learning process.**

Keywords*—Learning taxonomies; marzano and kendall taxonomy; personalization; XG boost; deep neural network; CNN; property graph; action verbs; content classification

I. Introduction

Learning is a process of adapting changes in personal and professional to ameliorate the quality of life. According to Stephen Hawking, Intelligence is the ability to adapt the change. Acquiring intelligence, absorbing, adapting and storing new information in memory is uneven among the learner. Hence it is the need of the hour to identify different learning characteristics of the learner to achieve a better learning outcome. The resource used to provide knowledge to the learner is known as learning content. According to the learner's preference and learning styles, learning content has to be provided to the learners. This process is called personalization in the learning process [1]. Personalized learning must pass some control over the learners, providing some input into how they progress through their learning activities. This can be achieved by adapting learning taxonomies in the learning process. Various taxonomies were developed by researchers in the field of Education and Learning since from the year 1956 [2].

This research work adapted Marzano and Kendall (MK) taxonomy to determine the learning style of the learners. MK taxonomy model provides better knowledge about certain fundamental processes in learning, such as emotion, memory, motivation and metacognition. This model also provides greater precision while creating learning objectives, having a more specific map of the types of knowledge that can be acquired and how they are acquired. Due to this greater precision, it is also possible to evaluate more easily [3]. MK taxonomy has six domains from lower order of thinking skills to higher order of thinking skills.

The investigator prepared the questionnaire based on the 2022 augmented verb list of MK taxonomy to find the learning style of the learner and classified the learners into six domains and 22 sub-domains of the considered taxonomy To classify the learning contents into micro contents the same taxonomy has to be utilized.

Text-based learning content was pre-processed to provide good interpretation and usage. It can also reduce the redundancy in the text content. After the content were pre-processed, it has to be classified based on the considered taxonomy. To accomplish the classification of learning contents into the micro-content process, the investigator has chosen seven algorithms from Machine Learning models such as

*Corresponding Author.

- Naïve Bayes,
- Support Vector Machine (SVM)
- K-Nearest Neighbor (KNN)
- Decision Trees
- Random Forest
- Bagging
- XG Boosting

and four algorithms from Deep Learning models such as

- Deep Neural Network (DNN)
- Recurrent Neural Network (RNN)
- Convolutional Neural Network (CNN)
- Recurrent Convolutional Neural Network (RCNN)

have been considered based on the verb list of six domains and sub-domains of MK's taxonomy.

The classified micro contents were stored in file format. And these micro contents can be represented using Property Graph also called a labelled property graph since it contains nodes(entities), edges(relationships) and properties(attributes). This research work creates ontology for MK Taxonomy to provide learning contents based on the weightage. In ontology, individuals are created for each micro-content with the annotation properties of learning content, keywords and file size. Once all the terms are arranged, the data can be retrieved using the SPARQL query. The representation in the property graph is visualized using the OWLGrEd Visualization tool.

Further, the researcher evaluated the performance of each model and compared them according to precision, recall, F1 score and accuracy. Classified micro contents obtained from the classifiers were mapped to the synonymous learners based on the maximum score on the accuracy of the model.

The rest of the paper is systematized as follows. Section II provides an overview of the related works. Section III provides the design and methodology of the proposed method. Section IV expresses evaluation and results and discussion. Section V illustrates the way to represent the learning content organization using a property graph and the method to extract the contents using SPARQL. Section VI discusses mapping the learning micro-content into the corresponding learner according to six domains and 22 sub-domains of the considered taxonomy. Section VII presents the conclusion and Section VIII illustrates the case study of the proposed method.

II. Related Works

A. Action Verbs

To express the noticeable behaviour of the learner the learning objective must start with action verbs. Action verbs were used to monitor the learner and the throughput of the learning objectives. Choosing the right verb for different types of the learner is an art [4]. The verb list of Marzano and Kendall Taxonomy was first published in 2007 and it needed an up-to-date update to include the later verbs. This is because the recent education system utilizes new vocabularies as per the current technology. The existing action verb list in the taxonomy may not be fulfilling to achieve the throughput of learning objectives. Hence it is the need of the hour to augment the verb list of Marzano and Kendall Taxonomy.

Augmentation is achieved by gathering suitable verbs from sixteen existing taxonomies and open domains. Hence the researcher has made an exhaustive search to update the verb list from 95 to 360 verbs as shown in Table I.

TABLE I. Augmentation of Verbs in MK Taxonomy

Domain / Level	Sub-domain/Level	No. of Existing Verbs	No. of Extended Verbs	Total Number of Verb List
Self-System Thinking	Examining Importance	01	15	16
	Examining Efficacy	01	14	15
	Examining Emotions	01	17	18
Metacognition	Examining Motivation	01	14	15
	Specifying Goals	02	13	15
	Process Monitoring	01	13	14
	Monitoring Clarity	01	06	07
	Monitoring Accuracy	02	12	14
Knowledge Utilization	Investigating	07	15	22
	Experimenting	05	14	19
	Problem Solving	06	10	16
	Decision making	04	11	15
Analysis	Specifying	04	11	15
	Generalize	05	06	11
	Analyzing errors	08	11	19
	Classifying	04	09	13
	Matching	08	10	18
Comprehension	Symbolizing	09	08	17
	Integrating	03	17	20
Retrieval	Executing	06	16	22
	Recalling	12	12	24
	Recognizing	04	11	15
	Total No. of Verbs	**95**	**265**	**360**

B. Adaptive Learning Path and Contents

A learning path is a progression of activities and concepts to be chosen by the learner to construct their knowledge or skills in a specific area. Traditional learning system provides the same content and learning path to all learners. But the learner's knowledge, circumstance, and preference are different, and their performance and satisfaction may decrease if they have been given the same content and learning path [5]. Presenting learner-centric learning content can ameliorate the effectiveness and performance of the learning process. To achieve this goal the researcher presented a new model as shown in Fig. 1.

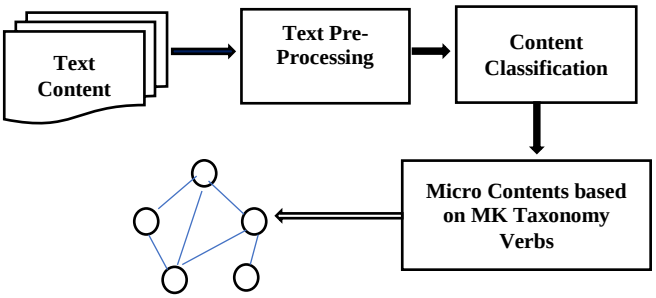


Fig. 1. A Text Content Classification Model.

These micro contents can be represented using Property Graph. Then the micro-content will be retrieved using SPARQL from the property Graph. Further, these micro contents are mapped into the corresponding learners according to six domains and sub-domains of MK taxonomy based on the descending order of the size of the micro contents.

C. Existing Content Classification Approaches

Learning Contents can be in the form of images, text, audio, animations, and video. Text contents can be classified as rule-based, supervised learning-based, and combined classifier-based approaches.

A set of handcrafted rules are utilized in the rule-based approach. In supervised learning text classification approach classification made based on learning past observations. The combined classifier utilized both a machine learning trained base classifier and a rule-based classifier for showing improvement in the throughput [6]. Table II illustrate the various researchers who proposed their model for the classification of questions into Bloom's taxonomy only on a cognitive level. This research work classifies the learning contents into six domains and 22 sub-domains of MK taxonomy.

TABLE II. VARIOUS STUDIES WERE CARRIED OUT TO CLASSIFY THE CONTENTS

S. No.	Name of the Researchers	Model Applied
1	Syahidah Sufi Haris et al [7]	Rule-Based Classification
2	Indika Perera et al [8]	Rule-Based Classification with n-gram Statistical Approach
3	Wen Chih Chang et al [9]	Rule-Based Classification with weighted Technique
4	Anbuselvan Sangodiah et al [10]	Support Vector Machine (SVM)
5	Anwar ali Yahya et al [11]	Support Vector Machine
6	Addin Osman et al [12]	Naive Bayes (NB), SVM, Logistic Regression, and Decision Tree
7	Norazah Yusof et al [13]	Artificial Neural Network
8	Dhuha Abdulhadi Abduljabbar et al [14]	SVM, NB and KNN use a majority voting algorithm.
9	Ali Danesh et al [15]	Combine three classifiers such as NB, KNN and Rocchio
10	Julio Villena Roman et al [16]	K- Nearest Neighbour

III. PROPOSED METHOD: DESIGN AND METHODOLOGY

Learning style is the strategy to accommodate receiving and processing the received information which are two phases of learning [17]. The process of recognizing the behaviour of the learner then spontaneously generates a natural learning path, and tailoring the learning contents to an individual learner is known as adaptation in learning which is the prime need for personalized learning [18]. Learning taxonomy can be employed to understand the learning levels of the learners scientifically. Hence this research focused on classifying the learning contents based on a learning taxonomy for better-personalized learning.

This research focused only on text learning contents. These learning contents were preprocessed and classified based on the augmented verb list of MK Taxonomy into micro contents. Then the suitable micro contents were assigned to the corresponding learner to accomplish the personalized teaching-learning process. Fig. 2 depicts the design architecture for text-based content classification. The design contains two main modules a pre-processing module and a Classification module.

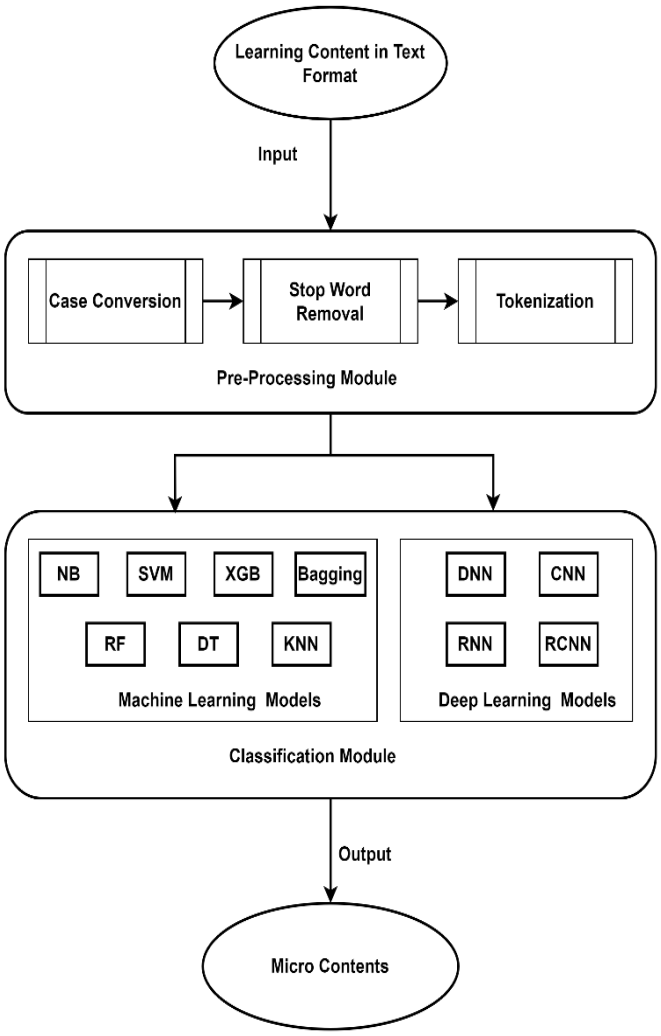


Fig. 2. Design Architecture of Content Classification.

The pre-processing module is the first module of content classification. Text pre-processing is essential for eliminating all the irrelevant objects from the data and making it ready for further processing. This is because raw text data might have insignificant text which makes it difficult to understand and investigate. Hence proper pre-processing must be implemented on raw text data [19]. This research work utilizes three pre-processing techniques as Case Conversion, Stop word removal, and Tokenization.

Case Conversion: Converting all the text content into the lower case is utilized to discard unproductive words [20].

Stop Word Removal: Articles, prepositions, pronouns, and conjunctions in any language are called stop words. "The", "a", "an", "so", and "what" are examples of stop words in English. Removal of such words would help in the size reduction of a dataset and further the training time can also be reduced due to the lesser number of tokens involved in the training [21].

Tokenization: Splitting text contents into smaller units is known as tokenization. The individual units are called tokens. Tokens can be words, phonemes, or maybe full sentences [22]. This research work utilizes sentence tokenization. The learning contents were divided into sentences and considered tokens.

In the e-learning environment, a large volume of learning materials was available in various formats. But it is necessary to provide appropriate learning content to the respective learners according to the six domains and sub-domains of MK taxonomy.

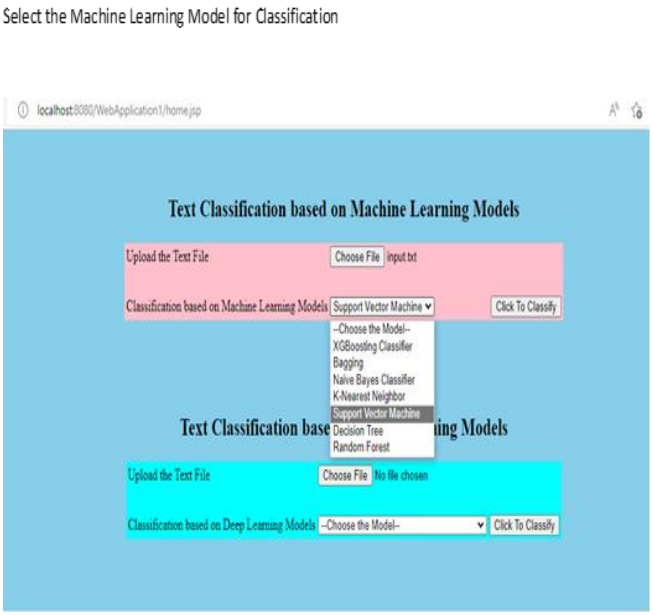


Fig. 3. Text Classification based on Machine Learning Models.

The classification module is the second important module of the design. The preprocessed text contents are classified using seven machine learning models Naive Bayes, SVM, Decision Trees, Random Forest, KNN, Bagging, XG Boosting as shown in Fig. 3, and four deep learning models such as DNN, CNN, RNN, RCNN algorithm based on the action verb list of 2022 augmented verb list of MK Taxonomy as shown in Fig. 4.

This study utilizes NetBeans IDE open-source integrated development environment using Java and libraries such as NLKT, pandas, TensorFlow, NumPy, sklearn, text blob, and seaborn for the classification process.

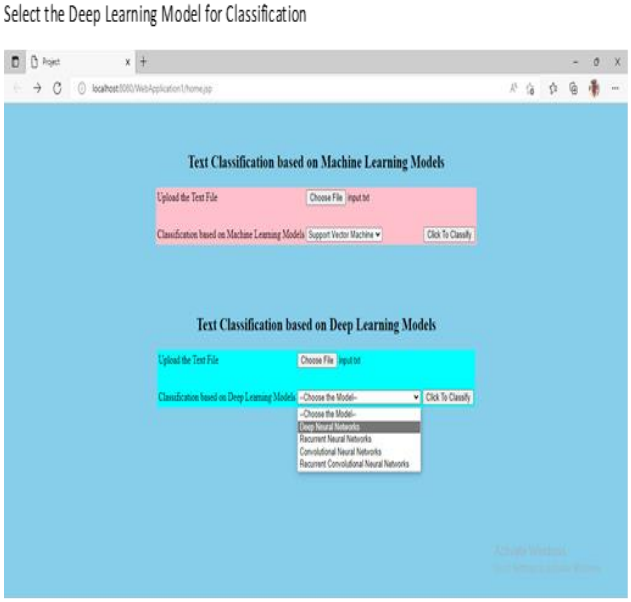


Fig. 4. Text Classification based on Deep Learning Models.

IV. Results and Discussions

For the evaluation process, this study has selected a dataset from the responses received from hundred students for three different learning contents. In a convenient sampling technique analysis on data-set can be carried out either by taking multiple sampling or by repeating the survey. The researcher adapted multiple sampling techniques to produce a reliable result.

A. Evaluation Metrics

Accuracy, precision, recall and f1-score are the measures for evaluation utilized by this study for understanding, measuring relevance and correctness of classification of learning content into micro-contents. Accuracy is used to check the correctness of the model. The exactness of the results is expressed by precision. The completeness of the quality of the results was measured by a recall. F1-score is the weighted average of precision and recall. F1-score is used to evaluate the binary classification system [23].

The evaluation of this study was performed based on the number of keywords classified per domain of MK Taxonomy. A maximum of ten keywords were considered for the classification of learning content into micro-content in each domain of the considered taxonomy.

B. Experiments

The experiments were conducted both on machine learning models and deep learning models and categorized into two.

Experiment 1: Analyze the results of individual Machine learning and deep learning classifier models.

Experiment 2: Results Analysis based on the evaluation metrics.

Experiment 1: Results of Individual Classifier Models

Table III represents the evaluation metric for the XG Boosting classifier in the machine learning model.

TABLE III. EVALUATION METRIC FOR XG BOOSTING CLASSIFIER

Marzano and Kendall Taxonomy Levels	No. of Keywords Classified per level	Accuracy	Precision	Recall	F1-Score
	1	68	0.69	0.7	0.69
Retrieval	2	69	0.7	0.68	0.69
Comprehension	3	71	0.64	0.72	0.68
Analysis	4	78	0.79	0.79	0.79
Knowledge Utilization	5	74	0.83	0.64	0.72
Meta Cognition	6	83	0.81	0.84	0.82
Self -System Thinking	7	82	0.84	0.75	0.79
	8	88	0.9	0.86	0.88
	9	88	0.9	0.85	0.88
	10	92	0.93	0.86	0.9
	Avg.	**80**	**0.8**	**0.77**	**0.78**

The overall accuracy of this classifier is 80%. The percentage of all the measures will be incremented if the number of keywords is increased in each level of MK Taxonomy.

Table IV represents the evaluation metric for the Bagging classifier in the machine learning model.

TABLE IV. EVALUATION METRIC FOR BAGGING CLASSIFIER

Marzano and Kendall Taxonomy Levels	No. of Keywords Classified per level	Accuracy	Precision	Recall	F1-Score
	1	65	0.57	0.74	0.65
Retrieval	2	59	0.60	0.56	0.58
Comprehension	3	59	0.62	0.54	0.58
Analysis	4	56	0.54	0.57	0.55
Knowledge Utilization	5	59	0.63	0.54	0.58
Meta Cognition	6	66	0.68	0.62	0.65
Self -System Thinking	7	53	0.55	0.46	0.50
	8	75	0.77	0.67	0.72
	9	81	0.79	0.82	0.80
	10	77	0.74	0.77	0.76
	Avg.	**65**	**0.65**	**0.63**	**0.64**

TABLE V. EVALUATION METRIC FOR NAÏVE BAYES CLASSIFIER

Marzano and Kendall Taxonomy Levels	No. of Keywords Classified per level	Accuracy	Precision	Recall	F1-Score
	1	71	0.81	0.65	0.63
Retrieval	2	74	0.82	0.65	0.72
Comprehension	3	75	0.67	0.78	0.73
Analysis	4	78	0.86	0.77	0.72
Knowledge Utilization	5	83	0.89	0.75	0.81
Meta Cognition	6	84	0.93	0.69	0.82
Self -System Thinking	7	86	0.85	0.92	0.80
	8	91	0.94	0.93	0.88
	9	91	0.92	0.90	0.93
	10	93	0.89	0.97	0.91
	Avg.	**83**	**0.86**	**0.80**	**0.80**

Table V represents the evaluation metric for the Naïve Bayes classifier in the machine learning model. This classifier achieved a considerable score in precision measurement. This shows the exactness of the results.

The evaluation metric for the SVM classifier is illustrated in Table VI. This study observed that the SVM classifier successfully classifies the content with much accuracy since the overall accuracy of the SVM classifier is 86%.

Tables VII, VIII and IX represent the evaluation metrics for KNN, Decision Trees and Random Forest classifiers.

Table X depicts the evaluation metrics for all the seven models in machine learning models.

TABLE VI. EVALUATION METRIC FOR SVM CLASSIFIER

Marzano and Kendall Taxonomy Levels	No. of Keywords Classified per level	Accuracy	Precision	Recall	F1-Score
	1	79	0.76	0.80	0.78
Retrieval	2	76	0.77	0.73	0.75
Comprehension	3	74	0.71	0.76	0.74
Analysis	4	85	0.84	0.86	0.85
Knowledge Utilization	5	84	0.87	0.76	0.81
Meta Cognition	6	86	0.83	0.91	0.87
Self -System Thinking	7	91	0.92	0.91	0.91
	8	95	0.95	0.95	0.95
	9	93	0.92	0.95	0.93
	10	97	0.96	0.98	0.97
	Avg.	**86**	**0.85**	**0.86**	**0.86**

TABLE VII. EVALUATION METRIC FOR KNN CLASSIFIER

Marzano and Kendall Taxonomy Levels	No. of Keywords Classified per level	Accuracy	Precision	Recall	F1-Score
	1	56	0.50	0.61	0.55
Retrieval	2	57	0.56	0.57	0.57
Comprehension	3	56	0.53	0.58	0.56
Analysis	4	58	0.59	0.56	0.57
Knowledge Utilization	5	66	0.69	0.60	0.64
Meta Cognition	6	53	0.58	0.45	0.51
Self -System Thinking	7	73	0.75	0.69	0.72
	8	83	0.84	0.81	0.82
	9	75	0.77	0.72	0.74
	10	85	0.85	0.84	0.84
	Avg.	**67**	**0.67**	**0.64**	**0.65**

TABLE VIII. EVALUATION METRIC FOR DECISION TREE CLASSIFIER

Marzano and Kendall Taxonomy Levels	No. of Keywords Classified per level	Accuracy	Precision	Recall	F1-Score
	1	42	0.42	0.42	0.42
Retrieval	2	55	0.51	0.56	0.53
Comprehension	3	43	0.46	0.42	0.44
Analysis	4	55	0.50	0.56	0.53
Knowledge Utilization	5	49	0.50	0.47	0.44
Meta Cognition	6	64	0.66	0.73	0.53
Self -System Thinking	7	58	0.60	0.59	0.48
	8	68	0.66	0.72	0.69
	9	56	0.53	0.55	0.59
	10	67	0.68	0.66	0.69
	Avg.	**56**	**0.55**	**0.57**	**0.53**

TABLE IX. EVALUATION METRIC FOR RANDOM FOREST CLASSIFIER

Marzano and Kendall Taxonomy Levels	No. of Keywords Classified per level	Accuracy	Precision	Recall	F1-Score
	1	66	0.56	0.69	0.62
Retrieval	2	74	0.67	0.78	0.72
Comprehension	3	67	0.67	0.67	0.67
Analysis	4	75	0.71	0.78	0.74
Knowledge Utilization	5	74	0.78	0.68	0.73
Meta Cognition	6	86	0.74	0.92	0.82
Self -System Thinking	7	80	0.81	0.79	0.80
	8	90	0.90	0.89	0.90
	9	85	0.80	0.89	0.84
	10	92	0.90	0.93	0.91
	Avg.	**79**	**0.75**	**0.80**	**0.78**

TABLE X. CONSOLIDATION OF THE EVALUATION METRICS FOR MACHINE LEARNING MODELS

Machine Learning Models	Accuracy (%)	Precision	Recall	F1-Score
XG Boosting	80	0.80	0.77	0.78
Bagging	65	0.65	0.63	0.64
Naïve Bayes (NB)	83	0.86	0.80	0.80
K-Nearest Neighbor	67	0.67	0.64	0.65
Support Vector Machine (SVM)	86	0.85	0.86	0.86
Decision Tree (DT)	56	0.55	0.57	0.53

The evaluation metric for Deep Neural Network is illustrated in Table XI. It is observed that the DNN classifier successfully classified the contents because the accuracy of the classifier is 83% which is high score than the remaining classifiers considered in this study.

The evaluation metrics for CNN and RNN were represented in Tables XII and XIII.

TABLE XI. EVALUATION METRIC FOR DEEP NEURAL NETWORK (DNN) CLASSIFIER IN DEEP LEARNING MODEL

Marzano and Kendall Taxonomy Levels	No. of Keywords Classified per level	Accuracy	Precision	Recall	F1-Score
	1	76	0.67	0.80	0.73
Retrieval	2	72	0.82	0.63	0.71
Comprehension	3	74	0.76	0.69	0.72
Analysis	4	76	0.65	0.86	0.74
Knowledge Utilization	5	78	0.84	0.75	0.79
Meta Cognition	6	85	0.82	0.87	0.84
Self -System Thinking	7	87	0.86	0.90	0.88
	8	93	0.95	0.91	0.93
	9	92	0.91	0.92	0.92
	10	96	0.98	0.92	0.95
	Avg.	**83**	**0.83**	**0.83**	**0.82**

TABLE XII. EVALUATION METRIC FOR RECURRENT NEURAL NETWORK (RNN) CLASSIFIER

Marzano and Kendall Taxonomy Levels	No. of Keywords Classified per level	Accuracy	Precision	Recall	F1-Score
	1	71	0.72	0.68	0.71
Retrieval	2	68	0.76	0.62	0.69
Comprehension	3	61	0.67	0.58	0.62
Analysis	4	68	0.68	0.67	0.68
Knowledge Utilization	5	74	0.75	0.73	0.74
Meta Cognition	6	79	0.82	0.74	0.78
Self -System Thinking	7	83	0.83	0.83	0.83
	8	86	0.81	0.91	0.86
	9	91	0.92	0.91	0.91
	10	92	0.91	0.94	0.93
	Avg.	**78**	**0.79**	**0.76**	**0.78**

TABLE XIII. EVALUATION METRIC FOR CONVOLUTIONAL NEURAL NETWORK (CNN) CLASSIFIER

Marzano and Kendall Taxonomy Levels	No. of Keywords Classified per level	Accuracy	Precision	Recall	F1-Score
	1	71	0.63	0.74	0.68
Retrieval	2	66	0.74	0.54	0.62
Comprehension	3	68	0.49	0.76	0.60
Analysis	4	64	0.60	0.70	0.64
Knowledge Utilization	5	68	0.79	0.57	0.66
Meta Cognition	6	74	0.73	0.76	0.74
Self -System Thinking	7	79	0.83	0.74	0.78
	8	87	0.86	0.88	0.87
	9	87	0.95	0.78	0.86
	10	93	0.93	0.93	0.93
	Avg.	**76**	**0.76**	**0.74**	**0.74**

The evaluation metric for the Recurrent Convolutional Deep Neural Network (RCNN) Classifier is illustrated in Table XIV. The combination of RNN and CNN is known as an RCNN classifier. The performance metrics of this classifier range from 75% to 78%.

Table XV provides the consolidation of the evaluation metrics for the four deep learning models.

Experiment 2: Overall Result Analysis based on the measures for evaluation.

The overall score obtained by all the classifiers using both machine learning and deep learning models were illustrated in Table XVI.

TABLE XIV. EVALUATION METRIC FOR RECURRENT CONVOLUTIONAL NEURAL NETWORK (RCNN) CLASSIFIER

Marzano and Kendall Taxonomy Levels	No. of Keywords Classified per level	Accuracy	Precision	Recall	F1-Score
	1	72	0.81	0.73	0.58
Retrieval	2	74	0.64	0.83	0.71
Comprehension	3	69	0.77	0.64	0.61
Analysis	4	69	0.84	0.57	0.64
Knowledge Utilization	5	66	0.45	0.78	0.62
Meta Cognition	6	69	0.55	0.69	0.83
Self -System Thinking	7	85	0.87	0.77	0.87
	8	86	0.85	0.79	0.91
	9	95	0.98	0.89	0.96
	10	87	0.93	0.84	0.82
	Avg.	**78**	**0.77**	**0.75**	**0.76**

TABLE XV. CONSOLIDATION OF THE EVALUATION METRICS FOR DEEP LEARNING MODELS

Deep Learning Models	Accuracy (%)	Precision	Recall	F1-Score
DNN	83	0.83	0.83	0.82
RNN	78	0.79	0.76	0.78
CNN	76	0.76	0.74	0.74
RCNN	78	0.77	0.75	0.76

TABLE XVI. CONSOLIDATION OF THE EVALUATION METRICS FOR ALL THE MODELS USED

ML and DL Models	Accuracy %	Precision	Recall	F1-Score
XGB	80	0.8	0.77	0.78
Bagging	65	0.65	0.63	0.64
NB	83	0.86	0.8	0.8
KNN	67	0.67	0.64	0.65
SVM	86	0.85	0.86	0.86
DT	56	0.55	0.57	0.53
RF	79	0.75	0.8	0.78
DNN	83	0.83	0.83	0.82
RNN	78	0.79	0.76	0.78
CNN	76	0.76	0.74	0.74
RCNN	78	0.77	0.75	0.76

Table XVII shows the Accuracy measure values obtained for each classifier and arranged in descending order based on the percentage of Accuracy.

According to the results, the SVM classifier performed well toward the correctness of classification and the accuracy is measured as 86 per cent. Further, the F1-score, the weighted

average of precision and recall is also 86% as shown in Table XVIII. The analysis based on accuracy is depicted in Fig. 5.

TABLE XVII. EXPERIMENTAL RESULTS AS PER ACCURACY

Machine Learning and Deep Learning Models	Accuracy %
Support Vector Machine (SVM)	**86**
Naïve Bayes (NB)	83
Deep Neural Networks	83
XG Boosting	80
Random Forest (RF)	79
Recurrent Neural Networks (RNN)	78
Recurrent Convolutional Neural Networks (RCNN)	78
Convolutional Neural Networks (CNN)	76
K-Nearest Neighbor	67
Bagging	65
Decision Tree (DT)	56

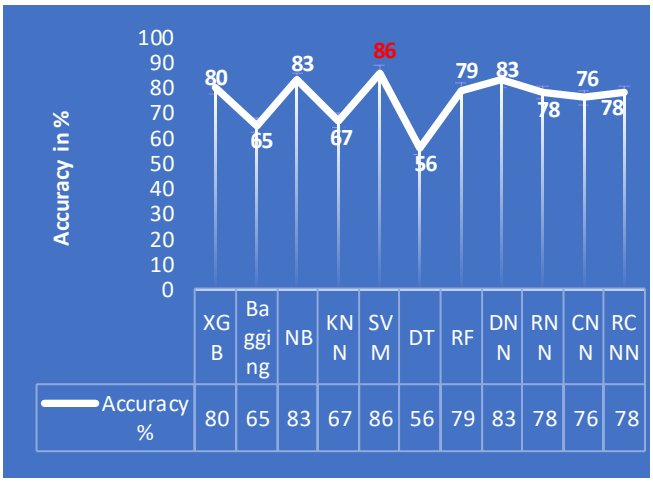


Fig. 5. Analysis based on Accuracy.

TABLE XVIII. EXPERIMENTAL RESULTS AS PER F1-SCORE

Machine Learning and Deep Learning Models	F1-Score
Support Vector Machine (SVM)	0.86
Deep Neural Networks	0.82
Naïve Bayes (NB)	0.8
XG Boosting	0.78
Random Forest (RF)	0.78
Recurrent Neural Networks (RNN)	0.78
Recurrent Convolutional Neural Networks (RCNN)	0.76
Convolutional Neural Networks (CNN)	0.74
K-Nearest Neighbor	0.65
Bagging	0.64
Decision Tree (DT)	0.53

As per the analysis, it is observed that the Naïve Bayes classifier achieved a considerable value in precision. The higher precision indicates that, less false positive measure. It shows the exactness of the classification of learning contents. Table XIX represents the experiment results based on the precision measure.

TABLE XIX. EXPERIMENTAL RESULTS AS PER PRECISION

Machine Learning and Deep Learning Models	Precision
Naïve Bayes (NB)	**0.86**
Support Vector Machine (SVM)	0.85
Deep Neural Networks	0.83
XG Boosting	0.8
Recurrent Neural Networks (RNN)	0.79
Recurrent Convolutional Neural Networks (RCNN)	0.77
Convolutional Neural Networks (CNN)	0.76
Random Forest (RF)	0.75
K-Nearest Neighbour	0.67
Bagging	0.65
Decision Tree (DT)	0.55

The completeness of the quality of the results was measured by a recall. SVM classifier again occupies the top place among other classifiers for the completeness of the classification of learning content according to keywords of MK Taxonomy. Table XX illustrate the experiment results based on recall measure.

TABLE XX. EXPERIMENTAL RESULTS AS PER RECALL

Machine Learning and Deep Learning Models	Recall
Support Vector Machine (SVM)	**0.86**
Deep Neural Networks	0.83
Naïve Bayes (NB)	0.8
Random Forest (RF)	0.8
XG Boosting	0.77
Recurrent Neural Networks (RNN)	0.76
Recurrent Convolutional Neural Networks (RCNN)	0.75
Convolutional Neural Networks (CNN)	0.74
K-Nearest Neighbor	0.64
Bagging	0.63
Decision Tree (DT)	0.57

According to the above analysis, this study concludes that the SVM classifier model provides more accuracy. Hence the micro-contents classified by utilizing the SVM classifier are considered for mapping to the synonymous learner based on the verb list of MK Taxonomy.

V. Learning Content Organization and Retrieval

Information can be represented in the form of tables, charts and graphs to support organizing, analyzing and fetching them when required. Graphs are shines well in representing the connections and relationships among diverse data.

A graph in which the connections between nodes are used to represent the relations along with name and some properties is called Property Graphs. Nodes, labels, relationships, and properties are the components of a property graph. Relationships and connecting nodes of data are capable of storing properties. Appropriate and easy-to-recognize labels are utilized by the property graph for modelling data and its connections. This structured form of data can be easily understood by laymen [24].

This research work utilizes the property graph to represent the classified micro contents according to MK taxonomy. Visualization of six domains and sub-domains and the corresponding micro-content of MK Taxonomy using OWLGrEd is depicted in Fig. 6. OWLGrEd is an editor for OWL to represent graphical notation in an ontology. Ontology becomes common in the fields of artificial intelligence and machine learning, where knowledge plays a vital role. Ontology creates a generic vocabulary which can be shared with researchers and different stack holders. It consists of machine-interoperable definitions of the domain concepts and the relationship between them. It enables the researcher to retrieve data based on knowledge which is known as knowledge-based retrieval. Ontology shares the knowledge to understand the structure of information which can be reused from the domain knowledge. This feature motivates to development of learning content to enhance the learner to gain knowledge of the subject based on their interest and learning ability.

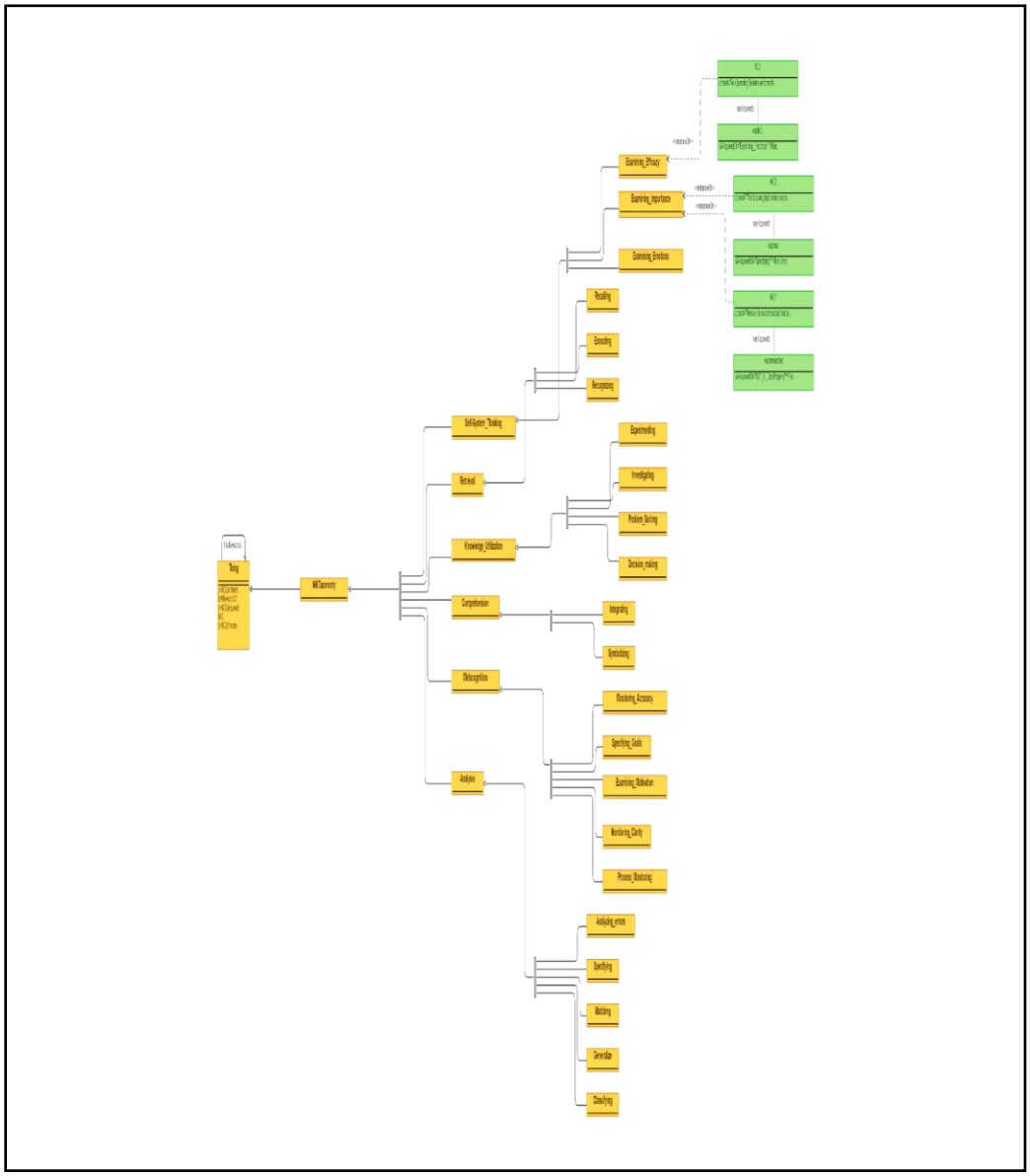


Fig. 6. Visualization of Six Domains and Sub-Domains and the Corresponding Micro-Content of MK Taxonomy using OWL.

Ontology is a collection of classes, properties, instances and axioms. Classes are also known as the concepts of the domain, properties define the relationship between the concepts, instances are the individuals of each class, and axioms denote the restrictions. Ontology can be defined as, a formal explicit specification of a shared conceptualization'. The key terms of a domain are identified and arranged hierarchically and the relationships between the terms are established before developing ontologies.

This research work creates ontology for MK Taxonomy to provide learning contents based on the weightage. The levels and sublevels of the considered taxonomy are arranged as classes hierarchically to frame ontology using the Protégé ontology development tool.

The level/domain of learning is identified through the keywords used in the learning content. Each level of MK taxonomy contains a different set of keywords to group the learning content. The keywords are listed as individuals and the relationship between the classes and keywords is established.

The learning contents were partitioned into micro contents to improve the learning ability of the learner. In ontology, individuals are created for each micro-content with the annotation properties of learning content, keywords and file size.

Micro-content (MC) can be represented as

$$MC_{ij} = \{K_{ij}, C_{ij}, FS(C_{ij})\} \quad (1)$$

Where

i represents domains of MK Taxonomy,

j represents sub-domains of MK Taxonomy,

K is a Keywords,

C is a Learning Content,

FS is the File Size of the learning content.

In this study, MC_{11} represents a micro-content in the sub-domain Recognizing in the domain Retrieval. Each micro-content is defined with these annotation properties to retrieve the content based on the file size given in Fig. 7.

Each micro-content is related to the type of class and object property it belongs. Variable content is created to hold the value of micro-content. Once all the terms are arranged, the data can be retrieved using the SPARQL query.

The SPARQL query to retrieve the micro-content based on the file size in descending order is given below and the result is shown in Fig. 8.

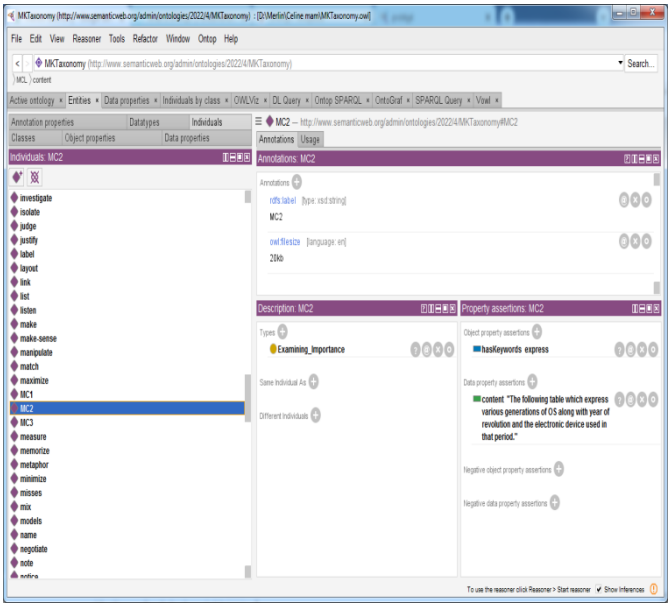


Fig. 7. Creation of Individuals for Micro-contents.

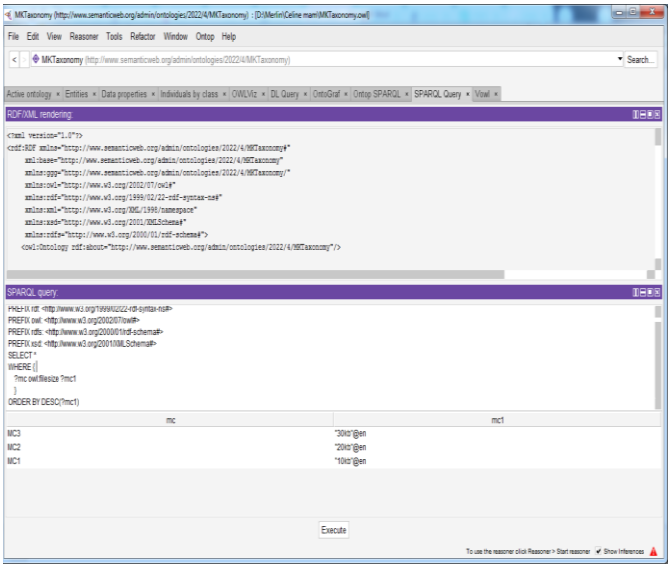


Fig. 8. Retrieval of the Micro-Contents using SPARQL.

SPARQL query to Retrieve the micro-content

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT *
WHERE {
 ?mc owl:filesize ?mc1
 }
ORDER BY DESC(?mc1)
```

VI. Mapping the Microlearning Content to the Synonymous Learners

The classified learning contents which are retrieved from the graph were mapped to the corresponding leaner to achieve personalization in the learning process. Each micro-content with learning content, keywords and file size defined with annotation properties was used to retrieve the content. Further, these retrieved micro-contents were arranged in descending order based on the size of the files. Based on the score obtained by the learner, they were classified by the researcher. The micro-content classified based on the keyword under the sub-domain Recognize in domain Retrieval is MC_{11}.

This study proposed a novel method to perform the mapping process. The learners' characteristics were obtained by the response received from them through the tool questionnaire according to the 2022 verb list of MK Taxonomy. Questions were rationalized to 50 according to six levels of MK Taxonomy as 8, 8, 10, 8, 8, 8 which can be considered as weightage (w) for each domain as shown in Fig. 9. Eight questions in D1 in turn sub-dived into 3, 3 and 2.

Dataset has been constructed from the response fetched from the hundred learners. The correct response was represented as 1 and the incorrect response was represented as 0. Further, the total score against each domain was calculated as illustrated in Table XXI. This provides a way to quantify each type of learning style in the learner.

Based on the score (SC) obtained by the learner out of each domain and sub-domains of MK Taxonomy, the number of micro-contents (NMC) retrieved from a graph as per each domain, and file size are the parameters for providing micro contents to the synonymous learner. Equation (2) is utilized for mapping the micro-contents to the corresponding learner.

$$K_{ij} = SC_{ij} / w(SD_{ij}) * NMC_{ij} \quad (2)$$

Where i represent six domains of MK Taxonomy,

j represents sub-domains of MK Taxonomy,

K - Number of micro-contents to be provided to the learner,

SC - Score obtained by the learner,

SD – Sub-Domains of MK Taxonomy,

w – Weightage assigned to SDs as shown in Fig. 9,

NMC – Number of Micro-Contents.

The Pseudo code for the mapping process is illustrated below.

1. Start the process.
2. If (i = 1) then j = 1 to 3
3. {Calculate NMCs for D1
4. $K_{ij} = SC_{ij} / w(SD_{ij}) * NMC_{ij}$
5. }
6. If (i = 2) then j = 1 to 2
7. {Calculate NMCs for D2
8. $K_{ij} = SC_{ij} / w(SD_{ij}) * NMC_{ij}$
9. }
10. If (i = 3) then j = 1 to 5
11. {Calculate NMCs for D3
12. $K_{ij} = SC_{ij} / w(SD_{ij}) * NMC_{ij}$
13. }
14. If (i = 4) then j = 1 to 4
15. { Calculate NMCs for D4
16. $K_{ij} = SC_{ij} / w(SD_{ij}) * NMC_{ij}$
17. }
18. If (i = 5) then j = 1 to 5
19. { Calculate NMCs for D5
20. $K_{ij} = SC_{ij} / w(SD_{ij}) * NMC_{ij}$
21. }
22. If (i = 6) then j = 1 to 3
23. { Calculate N MCs for D6
24. }
25. Stop the process.

The MCs were arranged in descending order based on the file size. Hence as per the above calculation, the MCs were mapped to the synonymous learners to achieve personalization in the learning process according to MK Taxonomy.

VII. Conclusion

The main objective of this paper is to specifically classify the learning contents based on the specific characteristics of the learner and according to the domains as well as the subdomains of the considered taxonomy. The learning contents in text format were represented in a property graph and retrieval of the same is achieved to fulfil the personalization process in the learner-centric environment. The learners were classified according to MK Taxonomy. Hence the classified learning contents were assigned to the synonymous learners to achieve personalization in the learning process.

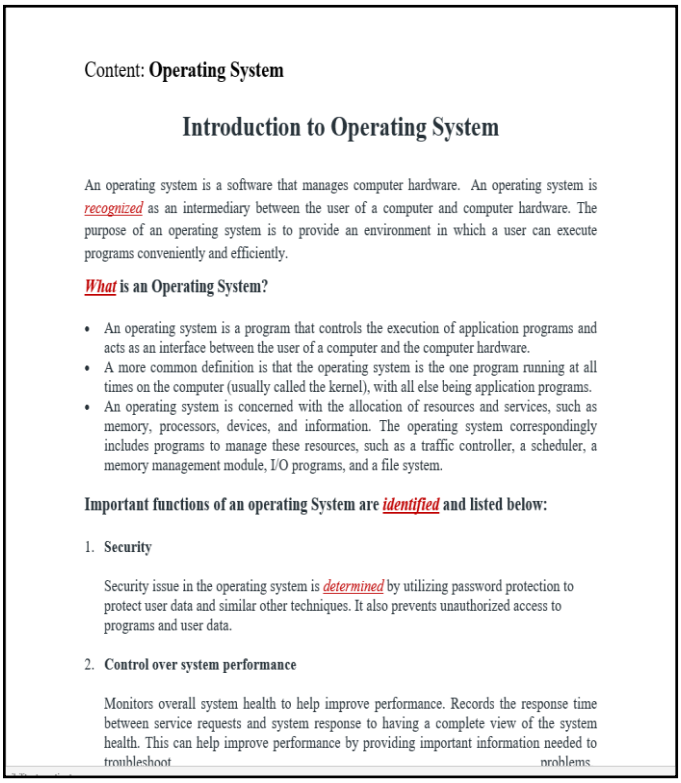
Many researchers classified the learners based on Bloom's Taxonomy's cognitive level. But this research work proposed a novel contribution towards the classification of learning contents into micro contents according to the six domains and 22 sub-domains of MK Taxonomy and represents them using a property graph. Further, these micro contents were retrieved from the graph and mapped to the corresponding learners who were classified according to MK Taxonomy. Hence the learner-centric learning contents were provided to the learners for better learning outcomes.

VIII. Case Study

Learning Contents classification can be carried out by the following steps. Fig. 9 shows the Screenshot of the learning content.

Input: Subject: Operating Systems-Tutor defined Text Contents.

An Operating System is recognized as an intermediate between the user of the computer and computer hardware. Important functions of an operating system are identified and listed below.



Content: **Operating System**

Introduction to Operating System

An operating system is a software that manages computer hardware. An operating system is *recognized* as an intermediary between the user of a computer and computer hardware. The purpose of an operating system is to provide an environment in which a user can execute programs conveniently and efficiently.

***What* is an Operating System?**

- An operating system is a program that controls the execution of application programs and acts as an interface between the user of a computer and the computer hardware.
- A more common definition is that the operating system is the one program running at all times on the computer (usually called the kernel), with all else being application programs.
- An operating system is concerned with the allocation of resources and services, such as memory, processors, devices, and information. The operating system correspondingly includes programs to manage these resources, such as a traffic controller, a scheduler, a memory management module, I/O programs, and a file system.

Important functions of an operating System are *identified* and listed below:

1. **Security**

   Security issue in the operating system is *determined* by utilizing password protection to protect user data and similar other techniques. It also prevents unauthorized access to programs and user data.

2. **Control over system performance**

   Monitors overall system health to help improve performance. Records the response time between service requests and system response to having a complete view of the system health. This can help improve performance by providing important information needed to troubleshoot problems

Fig. 9. Screenshot of Learning Contents.

Step 1: Text content Pre-Processing

Step 1.1: Case Conversion – convert into a lower case

an operating system is recognized as an intermediate between the user of the computer and computer hardware. important functions of an operating system are identified and listed below.

Step1.2: Stop word Removal

the operating system recognized intermediate user computer hardware. Important function operating system identifies list.

Step 1.3: Tokenization (Sentence Tokenization)

Token 1: operating system recognizes intermediate user computer hardware

Token 2: important function operating system identify the list.

Step 2: Verbs are Extracted from the tokens.

Verb list: recognize, identify, list

Step 3: Classification based on the verb list according to MK Taxonomy domains and sub-domains using seven ML models and four DL models as shown in Fig. 2. As per the performance metrics, the SVM model is used to classify this study. Keywords or the action verbs in MK Taxonomy were utilized for the classification of tokens into micro contents.

Action verbs 'recognize' and 'identify' the sub-domain Recognizing in domain Retrieval. Hence the corresponding MCs were assigned to that sub-domain.

Output:

MC 1: operating system recognizes intermediary user computer hardware.

MC 2: Important functions operating System identify list.

Step 4: These MCs were represented in the property graph as illustrated in Fig. 8 and retrieved using SPARQL.

Step 5: Mapping the MCs to the synonymous learners.

The total number of MCs in Sub-domain1 Recognizing in domain Retrieval were 02. These two MCs were to be mapped to the learners who were already classified under the same sub-domain as shown in Fig. 9 and the score obtained by the learners as shown in Table XXI were applied in the equation (2).

The score obtained by learner 1 in SD1 (SC) = 02

$$NMC = 02$$

$w(SD_{11}) = 03$

By utilizing equation (2) $K_{11} = 02/03*02 = 1.33 \approx 02$

Result:

Hence two MCs were provided to the learner in Sub-domain1 Recognizing in domain Retrieval according to MK Taxonomy in a personalized manner.

TABLE XXI. SCORE OBTAINED IN LEVEL 1 (RETRIEVAL) FOR FIFTEEN LEARNERS

Learner ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Score for Recognition	Score for Recalling	Score for Executing	Score for Leve 1-Retrieval
L1	**1**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**2**	**1**	**1**	**4**
L2	**0**	**1**	**0**	**0**	**1**	**0**	**1**	**0**	**1**	**1**	**1**	**3**
L3	**1**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**1**	**2**	**0**	**3**
L4	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**2**	**3**	**1**	**6**
L5	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**2**	**4**
L6	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**3**	**2**	**2**	**7**
L7	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**2**	**4**
L8	**0**	**0**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**2**	**1**	**4**
L9	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**1**	**1**	**1**	**2**	**4**
L10	**1**	**0**	**1**	**1**	**1**	**1**	**1**	**0**	**2**	**3**	**1**	**6**
L11	**1**	**0**	**1**	**0**	**1**	**0**	**0**	**1**	**2**	**1**	**1**	**4**
L12	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**1**	**2**	**1**	**1**	**4**
L13	**1**	**1**	**0**	**1**	**1**	**0**	**0**	**1**	**2**	**2**	**1**	**5**
L14	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**0**	**1**	**2**	**3**
L15	**1**	**0**	**1**	**1**	**0**	**1**	**1**	**0**	**2**	**2**	**1**	**5**

REFERENCES

[1] Sinem Aslana , Zehra Cataltepeb , Itai Dinerc , Onur Dundara , Asli A. Esmea , Ron Ferensc , Gila Kamhic , Ece Oktaya , Canan Soysala , Murat Yenera , " Learner Engagement Measurement and Classification in 1:1 Learning ", Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey.

[2] Cox, R.C. & Wildeman, C.E. (Eds.) (1970), " Taxonomy of Educational Objectives: Cognitive Domain; An Annotated Bibliography", Pittsburgh, PA: Learning and Research Centre.

[3] Marzano, R.J. (2001), " Designing a New Taxonomy of Educational Objectives", Thousand Oaks, CA: Corwin Press.

[4] https://study.com/academy/lesson/using-action-verbs-for-learning-objectives.html.

[5] Ahmad kardan, Maryam Bahojb Imani, Molood Ale Ebrahim, " A Novel Adaptive Learning Path Method", The 4th International Conference on e-Learning and e-Teaching, ICELET , February 2013.

[6] https://monkeylearn.com/text-classification/.

[7] Syahidah Sufi Haris and Nazlia Omar, " Determining Cognitive Category of Programming Question with Rule-based Approach", International Journal of Information Processing and Management, 4(3), 86-95, 2013.

[8] K. Jayakodi, M. Bandara, I. Perera, and D. Meedeniya. WordNet and cosine similarity based classifier of exam questions using bloom's taxonomy. International Journal of Emerging Technologies in Learning, 11(4):142–149, 2016.

[9] Wen Chih Chang and Ming Shun Chung. Automatic applying Bloom's taxonomy to classify and analysis the cognition level of english question items. 2009 Joint Conferences on Pervasive Computing, JCPC 2009, pages 727–733, 2009.

[10] Anbuselvan Sangodiah, Rohiza Ahmad, Wan Fatimah, and Wan Ahmad. A Review in Feature Extraction Approach in Question Classification Using Support Vector Machine. 2014 IEEE International Conference on Control System, Computing and Engineering, (November):536–541, 2014.

[11] AA Yahya and A Osman. Automatic classification of questions into Bloom's cognitive levels using support vector machines. In The International Arab Conference on December 2011, 2011.

[12] Anwar Ali Yahaya Addin Osman. Classifications Of Exam Questions Using Linguistically- Motivated Features : A Case Study Based On Bloom ' S Taxonomy Research Questions Research Aim. In The Sixth International Arab Conference on Quality Assurance in Higher Education, volume 2016, Saudi Arabia, 2016.

[13] Norazah Yusof and Chai Jing Hui. Determination of Bloom's Cognitive Level of Question Items using Artificial Neural Network. In 10th International Conference on Intelligent Systems Design and Applications (ISDA), pages 866–870, 2010.

[14] Dhuha Abdulhadi Abduljabbar and Nazlia Omar. Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination. Journal of Theoretical and Applied Information Technology, 78(3):447–455, 2015.

[15] Ali Danesh, Behzad Moshiri, and Omid Fatemi. Improve text classification accuracy based on classifier fusion methods. 2007 10th International Conference on Information Fusion, pages 1–6, 2007.

[16] Julio Villena Roman, Sonia Collada P ´ erez, Sara Lana Serrano, ´ and Jose Carlos Gonz ´ alez Crist ´ obal. Hybrid Approach Com- ´ bining Machine Learning and a Rule-Based Expert System for Text Categorization. In Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference — Twenty-Fourth International Florida Artificial Intelligence Research Society Conference — 18/05/2011 - 20/05/2011 — Palm Beach, Florida, EEUU, pages 323–328, 2011.

[17] Maria Dominic, Sagayaraj Francis, "An Adaptable E-Learning Architecture Based on Learners' Profiling", Published Online March 2015 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijmecs.2015.03.04 Copyright © 2015 MECS I.J. Modern Education and Computer Science, 2015, 3, 26-31.

[18] Bloom ,B. S., Engelhart .M. D., Furst E.J., Hill, W.H., & Krathwohl, D.R. (Eds.)(1956), " Taxonomy of Educational Objectives. The Classification of Educational Goals", Handbook I: Cognitive Domain. New York: David McKay Company, Inc.

[19] https://www.analyticsvidhya.com/blog/2021/08/why-must-text-data-be-pre-processed/.

[20] https://www.pluralsight.com/guides/importance-of-text-pre-processing.

[21] https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a.

[22] https://www.lexalytics.com/blog/text-analytics-functions-explained/.

[23] https://deepai.org/machine-learning-glossary-and-terms/f-score.

[24] https://www.dataversity.net/property-graphs-vs-knowledge-graphs/.

AUTHORS' PROFILE

First Author: Mrs S. Celine is currently working as an Assistant Professor in the Department of Computer Science, Government of Arts College for Men, Krishnagiri, Tamil Nadu. She has published four papers in International Journals. Her area of research is e-Learning using Deep Learning.

Second Author: Dr M. Maria Dominic is currently working as an Assistant Professor in Computer Science, at Sacred Heart College. He has published one book and more than 30 Research Articles in International Journals. His area of interest is AI, Machine Learning, and Deep Learning.

Third Author: Dr. F. Sagayaraj Francis is an Associate Professor in the Department of Computer Science and Engineering, Pondicherry Engineering College, Pondicherry, India. He has published more than 65 Research Articles in International Journals. He specializes in Data Management, Data Modeling, Information Systems and e-Learning.

Fourth Author: Dr M. Savitha Devi is currently working as an Assistant Professor and Head, Department of Computer Science, Periyar University Constituent College of Arts & Science, Harur, Tamil Nadu. She has published 30 research articles in various National and International Journals. She has published 2 books. Her area of research is Network Security.

The Hybrid Combinatorial Design-based Session Key Distribution Method for IoT Networks

Gundala Venkata Hindumathi[1], D. Lalitha Bhaskari[2]
Department of Computer Science & Engineering, JNTUK, Kakinada, India[1]
Department of Computer Science and Systems Engineering[2]
Andhra University College of Engineering (A),Andhra University, Visakhapatnam, India[2]

***Abstract*—Internet of Things (IoT) is currently being used in a range of applications as cutting-edge technology. IoT is a technological platform that connects the physical and digital worlds, allowing us to use things remotely. Various sensor-connected nodes serve as objects that communicate with one another over the internet. Hence security-related problems are more likely to arise in IoT networks. However, due to resource constraints such as power and memory capacity, complex security algorithms cannot be implemented in IoT networks. One of the security measures for IoT networks is to implement the lightweight key distribution algorithm. The lightweight key management process is essential for IoT networks to share the key securely. We presented the new key-distribution approach based on the hybrid combinatorial design that implements lightweight algorithms and describes the analysis functions. The comparison to existing hybrid combinatorial works shows better connectivity, resilience, and scalability.**

***Keywords*—Key distribution; hybrid combinatorial design; IoT networks; resource constraint nodes; symmetric key generation**

I. INTRODUCTION

The Internet of Things (IoT) is a system that allows multiple sensor nodes and wireless nodes to communicate without the need for human involvement. The term "things" in the Internet of Things refers to physical objects such as sensor nodes that monitor or access data from other networked devices. In the research aspect, IoT has been becoming a much-desired area. The security of each node's data is the primary issue in today's rapidly growing IoT networks.The security services are like confidentiality, authentication, and integrity of the data. Cryptographic algorithms and keys are required for encryption, and effective key management is essential for this process to work appropriately. Ineffective key management can make even the strong algorithms useless for any type of network. IoT networks also need to have strong key management procedures.

Even though key management is essential for IoT networks, using conventional key management methods demands more memory. Due to resource-constrained nodes' memory and battery limits, the IoT network requires a lightweight solution.Thus, we discussed about lightweight approaches that are already in use for key management.Basic methods to generate and distribute the keys to nodes in the network are symmetric keys and public keys. Even though the public key approach is widely used for key distribution, it could not be used often in IoT networks since it requires more memory and processing resources to run the code, and in many applications, these approaches are also costly. Hence, the Majority of IoT networks are using symmetric key distribution methods, which require only one key to share as mentioned by Alagheband et al. [1].

There are two methods for sharing keys amongst connected nodes: decentralized and centralized approaches. In the decentralized process, Nodes in the network can share their secret keys directly with one another to provide secure communication. Every node should hold private keys that are unique for communicating with each node in the network. Those private keys are exclusive to committed pairs only. However, as IoT networks grow, devices will be unable to keep as many secret keys in memory due to the restricted memory space of IoT nodes.

Another option for resolving this problem is to use a trustworthy centralized device to distribute private keys to all nodes in the network. Key Distribution Center (KDC) is an example of providing centralized service. Kouicem et al. [2] presented that the KDC is a mechanism that distributes keys to all the users in a network sharing sensitive or confidential information. When two nodes in a network need a connection, they request the KDC to generate a unique session key that end users can use as a secret key for communication. So, the nodes can share the data with other nodes connected to the network using Key Predistributions or KDC.

As a result, using a KDC with symmetric key distribution is the best way to distribute the key to all nodes. One of the best symmetric key generation approaches is combinatorial block designs.It uses a simple calculation to compute the blocks for different nodes. Many Authors have been working on this for determining the keys for multiple nodes.In the introduction, we covered the fundamental ideas of combinatorial block design, how the authors expanded these ideas to implement keys for every node, and a brief discussion on our approach.

Stinson et al.[3] used Balanced Incomplete Block Design (BIBD) which is one of the combinatorial designs to generate the blocks for sharing the keys securely with other nodes. When it is impossible to incorporate all treatments or factor combinations for every block, then BIBD is utilized here.

Assume there are b blocks, each with k keys, and v total number keys can be used, each key replicated r times. Thus,

$$br = vk$$

And also assume that the blocks (b) are just partially complete by confining with the following conditions.

1) $k < v$

2) In any block, the same key doesn't appear more than once.

$\lambda_{ij:}$ i and j are two different keys from the 'v', it gives the occurrences among the blocks.

Example 1: $v = 6, b = 4, k = 3, r = 2$

v={1,2,3,4,5,6}, b=no.of blocks, k=keys in each block,r= each key repitations in blocks.

So the Blocks are

b1:{1,2,3} ,b2:{1,4,5}, b3:{2,4,6}, b4:{3,5,6}

λ_{14}=1, λ_{46}=0 (It gives the pair occurrences in blocks.

In a Balanced Incomplete Block Design: $\lambda(v-1) = b(k-1)$.

Symmetric BIBD:

A BIBD is said to be Symmetric BIBD when $b = v; k = , r, \lambda = 1$

Example 2:

Consider (v,b,k,r,λ)=(7,7,3,3,1) because v=b;k=r

V(keys)={1,2,3,4,5,6,7}

b1:{1,2,3}

b2:{1,4,5}

b3:{1,6,7}

b4:{2,4,6}

b5:{2,5,7}

b6:{3,4,7}

b7:{3,5,6}

Another combinatorial method is the finite projection plane. A Finite Projection plane consists of P points and set of subsets of P called lines. A prime integer q (>=2) and that has four properties.

1) Every line should be having exactly q+1 points
2) Every point occurs on exactly q+1 lines
3) Exactly $q^2 + q + 1$ points used
4) Exactly $q^2 + q + 1$ lines used; then that can be called Symmetric Design with $(q^2 + q + 1, q + 1, 1)$ given by Stinson et al. [4].

Already existing key predistribution methods are mainly followed by three procedures.

1) Probabilistic: Keys are chosen randomly from the pool and assigned to the nodes.
2) Deterministic: Based on pre-defined procedures select the keys and assign them to the nodes.
3) Hybrid Approach: The combination of both approaches is mentioned above.

The KDC implements key predistribution methods to get the keys for all nodes. The pre-key distribution can be acquired based on the key-Matrix approach by Chien et al. [5], So it helped share the key easily. Other pre-key distribution approaches are Blundo et al. [6] and Liu et al. [7], In these, Polynomial-based key pre-distribution was proposed for group key establishment. In Chan et al. [8], Two nodes having q keys should be linked, and the hash value of the q keys would be used for key verification that improved resilience from the attackers. Qian and Sun [9] presented the drawback of the above approach is that resilience increased but wouldn't guarantee to get the common key between two devices. Li et al. [10] was provided threshold value for random key pre-distribution in which each should communicate with its neighbor node with the same key. Catakoglu et al. [11] increased the resiliency of the previous system by adding numerous key rings.

Camtepe andYener [12], first time they presented the symmetric balanced incomplete design(SBIBD) for generating the keys for nodes in the network, however, the disadvantage is the scalability of the network with nodes. In comparison to prior techniques, Lee et al. [13] exhibited improved resilience.Ruj et al. [14] generated the pre-key distribution method using the partial BIBD technique, however, it did not share the keys with every node in the network. Ruj et al. [15], the same authors proposed a combinatorial strategy for improving BIBD and PBIBD resilience. Bechkit et al. [16] employed a new pre-key distribution design, a combinatorial-based way to determine the keys, which improved the scalability and connectivity.Bahrami et al. [17] presented great scalability of the network nodes by using residual key pre-distribution design for key pool generation.

Camtepe et al. [18] presented a combinatorial method for generating keys for network nodes that are connected. And they used SBIBD and GeneralizedQuadrangle (QD), which are the basictwo deterministic key pre-distribution designs. Complete connectivity between network nodes was the improvement of this algorithm. Also provided is the hybrid pre-key distribution method.Chakrabarti et al. [19] and Kavitha et al. [20] enhanced the scalability and connectivity of the previous approach.Dargahi et al. [21] enhanced the hybrid method to get the keys for almost all network nodes, but didn't get the exact number of keys to all network nodes. When compared to prior hybrid techniques, Akhbarifar et al. [22] used a hybrid strategy and provided improved connectivity and resilience. However, the unique keys were not generated for nodes in the network.

Despite the fact that combinatorial designs have been addressed extensively, not all linked network nodes are given the session keys. Every IoT network needs to be able to enable the construction of many nodes and should distribute a session and a unique key for every node.By supplying unique and dynamic keys for each node, we suggested a hybrid combinatorial method that resolves the problems discussed earlier. As a result, our system now supports network scalability.

Our entire method is detailed in a total of six sections:

- Secton 1 gives the introduction part of basic methods for Combinatorial block designs.
- Section 2 explains the existing hybrid approaches and their drawbacks in detail.
- Section 3 is our actual work to be implemented to generate the unique and session keys for every node.
- Section 4 gives the analysis of scalability, connectivity,resilience, and Memory utilization. And also provides the results analysis with graphs.
- Section 5 is a complete discussion.
- Section 6 is a conclusion.

II. Related Works on Hybrid Combinatorial Designs

Camtepe and Yener [12] proposed a first-time pre-key distribution strategy based on the SBIBD technique. It was the basic combinatorial design to get the keys for network nodes.

Assume there are b blocks, each with k keys, and v total number keys can be used, each key replicated r times.The following criteria were used to allocate keys to the nodes in the proposed algorithm.

$q^2 + q + 1 = length\ of\ keypool(v); here\ q\ is\ prime$

$q^2 + q + 1 = number\ blocks(b); here\ q\ is\ prime$

$q + 1 =\ keys\ assigned\ to\ the\ each\ block(k)$

$q + 1 =\ In\ the\ blocks, every\ key\ is\ repeated\ (r)$

The fundamental advantage of this approach is that it identifies the unique keys among the b nodes. This technique had good connectivity and resilience, but it lacks scalability. However, this strategy has the disadvantage of limiting the total number of blocks that meet the before-mentioned criteria. As a result, it was completely reliant on the q value. This approach could not identify the keys for all n nodes in the network; where N is the total number of network nodes, and that was not meet the above condition.Although this method cannot be applied to all of the network's nodes, it accurately delivers the keys for the limited number of nodes.

In Camtepe et al. [18] (HSYM), the previous approach was upgraded by including scalability and resilience properties. It was implemented using a hybrid technique that enhanced the number of nodes in the IoT networks. It could find the b blocks by using SBIBD and this method found the complimentary design for all symmetric blocks then chose q+1 keys and assigns them to the remaining nodes.The author's implementation is described in Algorithm1.The fact that more nodes have a chance of acquiring the same key reduces the probability of obtaining a key share, which is a drawback of this technique.

Algorithm 1: Hybrid Design of HSYM

Input(s): N (Total Number of nodes)
Output(s): K (Block size)
Begin

1. Find largest prime power q such that $k \leq K$;
2. Generate base Symmetric
   - v objects $P = \{a_1, a_2, a_3, \ldots . a_v\}$
   - b blocks B= $\{B_1, B_2, B_3, \ldots . B_b\}$of size k;
3. Generate Complementary Design of the base design: Blocks $\bar{B} = \{\bar{B}_1, \bar{B}_2, \bar{B}_3, \ldots . , \bar{B}_b\}$where $\bar{B}_i = P - B_i$ and $|\bar{B}_i| = v - kfor\ 1 \leq i \leq b$;
4. Generate $N - b$ hybrid blocks $H = \{\bar{H}_1, \bar{H}_2, \bar{H}_3, \ldots . , \bar{H}_{N-b}\}$ of size *k*. For *i*th block $H_i$where $1 \leq i \leq N - b$:
   - Randomly select a block in $\bar{B}$, say $\bar{B}_j$
   - Randomly select a *k*-subset γ of the block B_j where $\gamma \notin H$,
   - Let $H_i = \gamma$ and $H = H \cup H_i$,
   - Use the variable $s_i$to hold index of the block $\bar{B}_j$from which the block $H_i$is obtained;
5. Blocks of the Hybrid Design are $B\ \cup H \Rightarrow K$

End

Dargahi et al. [21] (MHS) proposed an enhancement version of the above hybrid approach. For b blocks, they also used the same BIBD method. For the remaining nodes in IoT networks, they used a different key pool. N-b times, they have chosen q+1 keys from the new key pool that were assigned to additional N-b nodes.The generation of the blocks is described in Algorithm 2.The authors have used more space in the node memory to store the extra keys and new key pool in the nodes, but we all know, that IoT devices have limited capacity.

Algorithm 2: Hybrid Design of MHS

Input(s): N (Total Number of nodes)
Output(s): M Blocks
Begin

1. Find the largest prime number q Where $q^2 + q + 1 < N$
2. Generate the first symmetric $(q^2 + q + 1, q + 1, 1)$-BIBD with the following key pool:
   - $KP_1 = \{K_1, K_2, \ldots . , K_V\}$Containing *v* objects,
3. Generate *b* blocks $B = \{B_1, B_2, \ldots . , B_b\}$ from KP_1;
4. Choose a number *d* where $0 < d \leq q^2 + q + 1$;
5. Generate the second symmetric $(q^2 + q + 1, q + 1, 1)$-BIBD with the following key pool:
   - $KP_2 = \{K'_1, K'_2, K'_3 \ldots . , K'_V\}$ Containing *v* objects
   - KP_2 is generated in a way that *d* keys differ from KP_1 and other keys are the same,
6. Generate *b* blocks $M = \{M_1, M_2, \ldots . , M_b\}$ from KP_2;
7. Assign *b* blocks from *B* to *b* nodes $(b < N)$;
8. Choose$(N - b)$blocks from *M* in a random manner and assign them to $N - b$ remaining nodes

End

Akhbarifar et al. [22] (MHSYM) proposed a new enhancement of the previous proposes. They identified two random blocks in b, combined their blocks data, then extracted the q+1 keys from it. Then allocated each of the remaining nodes with a random selection of q+1 keys.Algorithm 3's detailed explanation of the entire process. The key share probability was increased as compared to Camptepe [18] method, but there is no guarantee that at least one common key would be allocated among the blocks. The complete procedure explained in Algorithm 3.

Algorithm 3: Hybrid Design of MHSYM

Input(s): N (Total Number of nodes)

Output(s): M Blocks

Begin

1. Find the largest prime number q
2. Where $q^2 + q + 1 < N$
3. Generate the first symmetric $(q^2 + q + 1, q + 1,1)$-BIBD with the following key pool:
   - $KP_1 = \{K_1, K_2, \ldots, K_V\}$containing *v* objects,
4. Generate b blocks $B = \{B_1, B_2, \ldots, B_b\}$ from $KP_1$and assign them to *b* nodes;
5. Choose two blocks among *b* blocks randomly;
6. Merging two blocks to construct new key-pool *M*;
7. Select $(N - b)$ blocks among $q + 1$ subsets of M and assign them to $N - b$ remaining nodes.

End

As mentioned above, the Procedures to apply the hybrid combinatorial design won't generate keys for all blocks of nodes.Our method uses a limited memory source to provide session keys for all blocks of nodes.The proposed work covers related algorithms and also provides examples for key generation.

III. Our Proposed Work

The Symmetric BIBD (SBIBD) allows multiple users in the same network to share the same keys without causing any problems. The IoT Architecture has not been supporting for huge capacity of memory inbuilt and high processing devices. Because of the above-mentioned reasons, the IoT node connected to the network is unable to remember all of the keys required for communication with other nodes in the network.

The SBIBD allows for the storage of the smallest amount of keys on the devices themselves, however, scalability is an issue here. If the network grows larger, nodes will be unable to store numerous keys in the tiny size memory. So, we are providing a new solution to this problem, a centralized system called Dynamic Key Generation and Distribution Center (DGDC). And the whole design that we suggest is depicted in Fig. 1.

In the context of IoT, we describe the symmetric key authentication and key management system based on BIBD. In this paper, we present a technique for exchanging the secret key that uses for providing the different security levels to assure scalability and confidentiality. We propose a technique for key agreement between two IoT devices that have never been in contact before, based on trusting the centralized server or using a proxy-based approach.

Fig. 1, describes the overall architecture that we have implemented to generate the session key and distribute it to the host which is requestedto the centralized server. The diagram itself is made up of three different blocks: DGDC, Initiate System (A) which starts to set up the communication connection, and Destination System (B) which accepts data from User (A) after receiving the Session key from DGDC.

The connected systems first exchanged their symmetric key to communicate with the centralized block, which is DGDC. Before implementing this architecture, the symmetric keys (secret keys for authentication) should be shared with DGDC so that other systems already connected to the network can communicate with it. Hence, this step is really important for our design because it is also providing authentication.Key generation and Key Distribution are the two main components of DGDC's actual work.

For Generating the keys, DGDC always works on the below-mentioned algorithms to implement the symmetric keys for all connected nodes. The previous algorithms mentioned in the related works are not implementing unique keys for all connected nodes.It is a pioneering building component for dynamic key implementation and distribution, increasing data security by often changing node keys.

In the DGDC, Data generation block contains all of the modules that have been proposed to create dynamic and unique keys for data transactions carried out by connected nodes. The modules are:

1) SBIBD,

2) Building the remaining nodes,

3) Computing the unique keys for each node in the network using a hybrid combinatorial design approach,

4) Reconstructing the outgoing blocks of nodes to protect the keys that have been compromised.

To create a complete table with unique keys for every node, DGDC executes each module in the order that they are presented.Once the table has been built, DGDC verifies requests using secret keys before sending the session key to the requested nodes.

Here, the architecture also proposed by us gives more security levels to the data because the session keys are not known by each individual connected system in the network. If an attacker compromises one of the systems, the attackers are unable to identify the session keys from the compromised system as it never stores any keys in their systems.

In particular, eight steps must be completed to observe the workings of our model. They are mentioned below in the Fig. 1. The model can generate and distribute the session key for communication between the request systems based on the mentioned processes. One of the most essential features of the proposed approach is the ability to dynamically alter the session keys of each system.

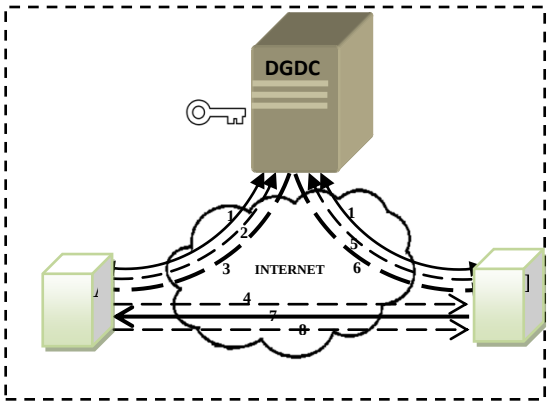


Fig. 1. Proposed Architecture.

1) Both users A and B identified their symmetric keys and exchanged them with DGDC for authentication purposes.

2) User A requests a session key from DGDC to communicate with User B.

3) By using a symmetric key, DGDC completes the authentication process and obtains the common key of both parties. And sends it to User A.

4) Using the Session key provided by the DGDC, User A transfers the data to User B.

5) Using its symmetric key, User B requests the session key from DGDC.

6) Like Step3, DGDC finishes its authentication process and provides the session key (which is already shared with A) to B.

7) User B uses the session key to decrypt the data provided by User A and provides the acknowledgment in an encrypted format.

8) The communication between User A and User B begins with the use of the same session key.

In the Proposed Work, The main required module is DGDC, it is generating the session keys dynamically and distributes them to the systems. First, we have to complete the code for generating the SBIBD with restricted blocks provided in Algorithm 4. The session keys for all nodes in the network could not be generated through the SBIBD procedure.

Where N is the number of network nodes used in communication. Calculate N>=q2+q+1, where q is the largest prime integer that may be used to solve the preceding equation; the result is v and b. Here, The 'N' and the 'v ' may not be the same. That is, the SBIBD algorithm was unable to determinethe keys for each node in the N network. SBIBD can be generated with v blocks and q+1 keys, which are represented by the k in each block.

The input for Algorithm 4 is N which is the number of nodes that need to be connected to the network, where v, k, and r are generated by the above Algorithm 4. The maximum number of nodes (blocks) in a network for generating session keys in SBIBD is represented by b. However, Algorithm 4 provides limited session keys for a few numbers of network nodes, therefore we are improvising by using other Algorithms 5, 6, and 7.

Algorithm 4: Design of SBIBD

Input(s): N (Total Number of nodes)
Output(s): B
Begin

1. Choose the maximum prime number q to compute the below equation
$$q^2 + q + 1 \leq N$$
2. Using the previous equation, generate inputs for producing the blocks.
$v = q^2 + q + 1$; where v is the size of the key pool
$b = q^2 + q + 1$; b is the number of blocks
$k = q + 1$; k is the number of keys allotted to each block
$\gamma = 1$; γ denotes, In SBIBD, each node has only one shared key to communicate to other nodes in the B.
3. Construct blocks B using Symmetric BIBD design
$$B = SBIBD(v, b, k, \gamma).$$
Then assign the blocks in $B = \{B_1, B_2, .., B_b\}$

End

Algorithm 5 completes the generation of remaining blocks of the network nodes. Algorithm4 computes the ' B ' number of blocks, while Algorithm5 will handle the rest.c=N-b; c is the number of blocks to be calculated, where N network nodes and b have already been given in Algorithm 4. Algorithm 5 determines which of the c number blocks should be assigned to the network's other nodes. In Algorithm 5, the R represents the remaining nodes of the IoT network.Select the keys from the key pool, and then place them as keys to generate the blocks by the requirements of Algorithm 5.

Algorithm 5: Design for remaining nodes(R)

Input(s): c,v,k
Output(s): R
Begin

1. Construct the (v,N-b,k,r,γ); here v is the key pool, N-b blocks need to construct, k keys for each node, r repetitions among the blocks, γ = 2 or more; means each block in N-b should share two or more keys among the q+1 keys.
2. As a result, each key from the key pool can only be used at most 3q times in the construction of N-b blocks.
3. Then return R blocks from this Algorithm
$$R = \{B_{N-b,}B_{N-b+1,} \dots, B_N\}$$

End

The final blocks are represented by $H = B \cup R$ which is input for Algorithm 6 and also computed the key pair values for all resource-constrained nodes.

Algorithm 6 is used to generate the v number of keys, however, the remaining keys were not able to be generated directly. The remaining c keys are found and perform an XOR operation on the common keys that existed between the two nodes. At the end of Algorithm6, be able to find the unique session keys between each node in the network. Here, Algorithm6 uses 32 bit (8 bytes) key for computation as the IoT devices could be handled easily with this length.

Algorithm 6: Hybrid Combinatorial Design with Unique keys

Input(s): N,v,b,K,B

Output(s): H(Total no.of blocks), x (The Session Key)

Begin

1. Execute Algorithm1 to get the q^2+q+1 Symmetric block within N blocks.

   $v = q^2 + q + 1$ (Number of keys used)

   $$Key = \{K_1, K_2, .., K_v\}$$

   $b = q^2 + q + 1$ (Number of nodes generated with the length of k by Algorithm1)

   $$B = \{B_1, B_2, .., B_b\}$$

2. Generate N-b blocks using Algorithm2.

   $$R = \{B_{N-b}, B_{N-b+1}, \dots, B_N\}$$

   $$Key = \{K_1, K_2, .., K_v\}$$

3. Hybrid Design's Blocks are $H = \mathrm{B} \cup R$
4. Choose any two blocks from N (BB, BR, and RR) blocks randomly and determine the common key(s) of these blocks that should store in *l*.

   Example: Here we have taken two blocks B_1, B_{N-b}.

   $$l = B_1 \cap B_{N-b}$$

   Get the common keys that are presented in both blocks.
5. (i) If the length of the *l* is one then directly take the key as the secret key for both blocks.

   $$if\ length(l) == 1\ ;$$

   $$then\ x = l\ and\ x\ as\ a\ secretkey$$

   (ii) If the length of the *l* is above one, take the last two keys from blocks and do the XOR operation among those keys.

   $$if length(l) \geq 2$$

   $$then x = l[length - 1] \oplus l[length - 2]$$

   $$x\ is\ a\ secretkey$$

   (iii) If the length of the *l* is null, select the first key-value from each block and calculate XOR between those keys. For example

   $$if\ length(l) == 0$$

   $$then\ x = first\ key(B_1) \oplus first\ key(B_{N-b})$$
6. *x* is the final secret key that is given by the DGDC.

End

The blocks for nodes are generated by DGDC up through Algorithm 6, and those key values in blocks aresent to nodes during transaction time. Once generated, they can be used every time, so there are chances of keys being compromised. Thus, We have also implemented an Algorithm 7 to get a solution for compromised keys by an attacker. Algorithm 7 illustrates how we can avoid attacks by utilizing a technique that shuffles the keys in the blocks in a certain amount of time.

Algorithm 7: Reconstruction of H

Input(s): H (Total blocks with keys)

Output(s): H blocks

Begin

1. For every, Threshold time(T) changes the key values of nodes
2. Shuffle all blocks of the H and assign the values of the block to nodes

   $$H = shuffle(H) \text{ for } \Delta T$$
3. And shuffle each block key value of the H to get the session key from Algorithm3.

   $$Apply\ \forall Bin(H);\ like$$

   $$H(B_i) = shuffle(H(B_i)) \text{ for } \Delta T$$

End

The complete workflow illustrates the DGDCs in Fig. 2.

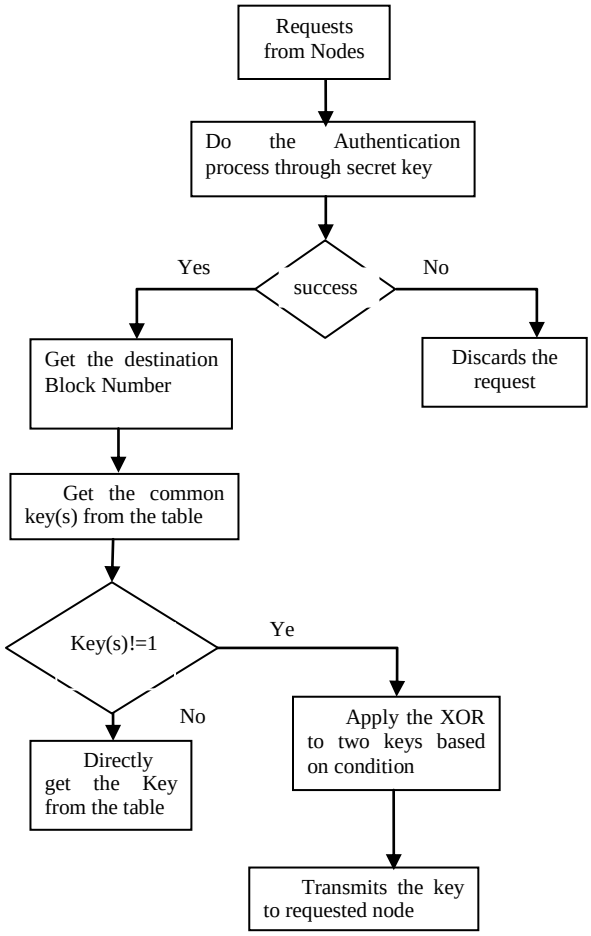


Fig. 2. The Workflow of Dynamic Key Generation and Distribution Center (DGDC).

Example 3: In the Example, We are taking network size as 7 (Select maximum prime number that satisfies q=2, 22+2+1=7). So.

v= 2^2+2+1= 7, k=q+1=4

Such that the total number of blocks (B) designed by the SBIBD=7. Each DGDC module identifies the session key for every block by using Algorithm 4. And below Table I shows the key numbers for each block.

These are the seven keys stored in DGDC before starting the communication.

{1: 31037803, 2: 34051950, 3: 75095512, 4: 67731601, 5: 90790958, 6: 42721930, 7: 56819008}

By using the above network configuration the users can communicate with each other. For example, if User 1(B1) wants to transmit the data to User 6 (B6), DGDC identifies the

common key between B1 and B6 i.e., 3. Then select the '3' Key value from Algorithm 3 that is already given above. Here the value is 75095512. DGDC transmits this Key to User 1 as well as User 6 when it sends a request for communication.

TABLE I. CONSTRUTION OF 7 NODES USING ALGORITHM1

User(s)	Block Number	Key number1	Key number2	Key number3
1	B1	1	2	3
2	B2	1	4	5
3	B3	1	6	7
4	B4	2	4	6
5	B5	2	5	7
6	B6	3	4	7
7	B7	3	5	6

Example 4: Here, We are taking 20 as the input, N (q=3, 32+3+1=13). We could not use the value for q is 2.

$v = q^2 + q + 1 = 13, k = q + 1 = 4$

These blocks are getting from the DGDC from Algorithm4. But the given N value is 20. So, we have to find out the other blocks by using Algorithm 5. Table II shows the key numbers up to block 13.

The below-mentioned keys are the basic keys that are stored in the DGDC and these keys are also used for calculating the other node keys by using Algorithm 6.

{1: 56940651, 2: 83179189, 3: 88850165, 4: 50901991, 5: 95809326, 6: 88046686, 7: 45506527 , 8: 42631960, 9: 36152950, 10: 31237906, 11: 91772959, 12: 87834612, 13: 13247806}

The Remaining nodes are: 20-13=7. Table III shows the key numbers of the remaining nodes.

TABLE II. CONSTRUCTION OF 13 NODES USING ALGORITHM1

User (s)	Block Number	Key Number1	Key Number2	Key Number3	Key number4
1	B1	1	2	3	4
2	B2	1	5	6	7
3	B3	1	8	9	10
4	B4	1	11	12	13
5	B5	2	5	8	11
6	B6	2	6	9	12
7	B7	2	7	10	13
8	B8	3	5	10	12
9	B9	3	6	8	13
10	B10	3	7	9	11
11	B11	4	5	9	13
12	B12	4	6	10	11
13	B13	4	7	8	12

TABLE III. CONSTRUCTION OF REMAINING 7 NODES USING ALGORITHM 2

User (s)	Block Number	Key number1	Key number2	Key number3	Key number4
14	B14	1	2	4	7
15	B15	1	2	4	10
16	B16	2	4	10	13
17	B17	2	4	9	10
18	B18	1	4	7	9
19	B19	4	7	10	13
20	B20	4	7	10	11

Algorithm 5 can generate multiple possibilities to build the tables to address the aforementioned problem.One of the solutions has mentioned in Table III. The DGDC can select any

But, here we can get the duplicate key numbers for identified blocks. We have implemented Algorithm 6 to calculate the accurate key for both parties. For Example, User 1 (B1) wants to send the data to User 17 (B17). So, DGDC needs to identify the key for them by using Algorithm6 itself.

The block key numbers are again mentioned here for reference.

B1-(1, 2, 3,4)

B17-(2,4,9,10)

Two common keys from the above blocks are 2 and 4. The keys values are taken from above dictionary for 2: 83179189 and 4: 50901991. After applying Algorithm 6,the output key-value is D3878818. So, DGDC transmits this common key to both users for further communication.

We shall receive new blocks for nodes after the same table with keys has been used for a time determined by the DGDC.

IV. ANALYSIS

The connectivity, scalability, resilience, and memory utilization of our model are all evaluated.

A. Scalability

The model can be scalable with the maximum number of nodes that were constructed for the IoT network. The model works with all keys in the keyring that correspond to the maximum number of IoT nodes that can be supported. The number of blocks generated with their keyrings determines the network's scalability. The scalability of a proposed approach is

$$q^2 + q + 1 + \binom{n^2 + 2qn + n}{q + 1}$$

here n is an integer value to get the next prime number and$(q^2 + q + 1)$ is identified by Algorithm1.

The following equation is for the calculation of the remaining nodes:

$(q + n)^2 + (q + n) + 1 - (q^2 + q + 1) =$

$n^2 + 2qn + n$

B. Connectivity

The probability of any two IoT nodes sharing only one communication key.

The main advantage of this model is to get the probability of key share at most 1 for maximum all cases.

$$p_{BB} = \frac{\binom{q^2+q+1}{2}}{\binom{N}{2}} = \frac{q^2+q+1(q^2+q)}{N(N-1)}$$

$$p_{BR} = \frac{\binom{q^2+q+1}{1}\binom{N-(q^2+q+1)}{1}}{\binom{N}{2}}$$

$$= \frac{2(q^2+q+1)\left(N-(q^2+q+1)\right)}{N(N-1)}$$

$$p_{RR} = \frac{\binom{N-(q^2+q+1)}{2}}{\binom{N}{2}}$$

$$= \frac{(N-(q^2+q+1)(N-(q^2+q+1)-1)}{N(N-1)}$$

According to the proposed model $p_{BB} + p_{BR} + p_{RR} = 1$, because the connectivity should be 1 in all maximum cases in the proposed approach.

$$p_{keyshare} = p_{BB} + p_{BR} + p_{RR} = 1$$

C. Resilience

Resilience means reliability among the network nodes from the attacker. The capture attack is called by capturing and revealing the key values from the nodes. So, the links which are used by the attacked key that might be compromised then those links are at risk. The proposed approach employs a unique key to communicate across nodes. And at random times, it shuffles all key values of blocks and blocks values as well As a result, if an attacker captures a key, it will not be worked after the shuffle.

$$p(L|C_x) = \sum_{\forall i} p(l_i|l)(p(D_i|C_x)$$

Where L denotes the link, $C_x$is x nodes are captured, l_i is the secure link between devices that already shared the i^{th} key in the pool. $D_i$has identified the key pool that includes key $i$is compromised. In our proposed system, from Algorithm 3, each key appears in the B blocks.

$r = q + 1$. For R blocks, each key repetitions are, $r' = 3q$

$$p(l_i|l) = \frac{\binom{((q+1)+3q)}{2}}{\binom{q^2+q+1+\binom{n^2+2qn+n}{q+1}}{2}}$$

The probability of key$i$, appearing in one or more of the x compromised keyrings is:

$$p(D_i|C_x) = 1 - \frac{\binom{((q^2+q+1)+(N-(q^2+q+1)))-((q+1)-3q)}{x}}{\binom{q^2+q+1+\binom{n^2+2qn+n}{q+1}}{x}}$$

When x keyrings are captured, the probability of a link being compromised can be calculated as.

$$p(L|C_x) = \sum_{i=1}^{q^2+q+1} p(l_i|l)(p(D_i|C_x)$$

$$= q^2+q+1\frac{\binom{((q+1)+3q)}{2}}{\binom{q^2+q+1+\binom{n^2+2qn+n}{q+1}}{2}} p(D_i|C_x) \cong p(D_i|C_x)$$

Our proposed system increases resilience when compared to previous models. The other systems probabilities of resilience are:

In the [18] model:$p(L|C_x) = 1 - \frac{\binom{q^2}{x}}{\binom{q^2+q+1}{x}}$

In the [21] model:

$$p(L|C_x) = 1 - \frac{\binom{2q^2}{x} + 2\binom{q^2}{x}}{\binom{2q^2+2q+2}{x}}$$

In the [22] model:

$$p(L|C_x) = 1 - \frac{\binom{((q^2+q+1)+(N-(q^2+q+1)))-((q+1)+\binom{2q}{q})}{x}}{\binom{q^2+q+1+\binom{2q+1}{q+1}}{x}}$$

Resilience values are provided for 500, 800, and 1700 nodes in Tables IV, V, and VI, respectively. Fig. 3, 4, and 5 show the graphs for the corresponding tables with various nodes. Different methods for hybrid combinatorial design are provided in tables and figures, and it is demonstrated that our approach produces the best results when compared to other ways.

TABLE IV. RESILIENCE VALUES FOR 500 NODES

N	Compromized nodes x	q value	HSYM	MHS	MHSYM	Our Approach
500	20	19	0.669	0.664	0.644	0.639
	40		0.897	0.89	0.87	0.852
	60		0.97	0.965	0.942	0.939
	80		0.992	0.98	0.975	0.971
	100		0.998	0.996	0.986	0.984
	120		0.999	0.999	0.992	0.99
	140		0.999	0.999	0.999	0.998

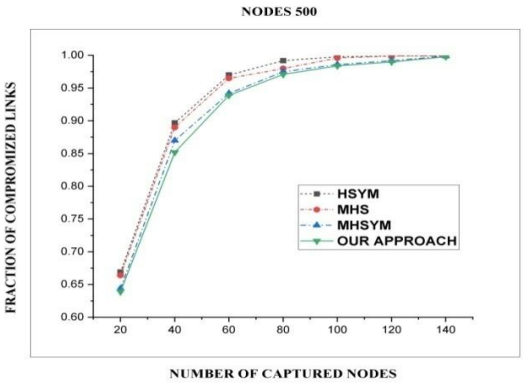


Fig. 3. Resilience Simulation Results of Our Approach Versus HSYM[18], MHS[21], and MHSYM[22] for the 500 Nodes.

TABLE V. RESILIENCE VALUES FOR 800 NODES

N	Compromized nodes x	q value	HSYM	MHS	MHSYM	Our Approach
800	40	23	0.84	0.83	0.81	0.8
	60		0.94	0.93	0.93	0.91
	80		0.97	0.97	0.96	0.95
	100		0.994	0.99	0.99	0.98
	120		0.997	0.996	0.99	0.99
	140		0.999	0.998	0.99	0.99
	160		0.999	0.999	0.99	0.99

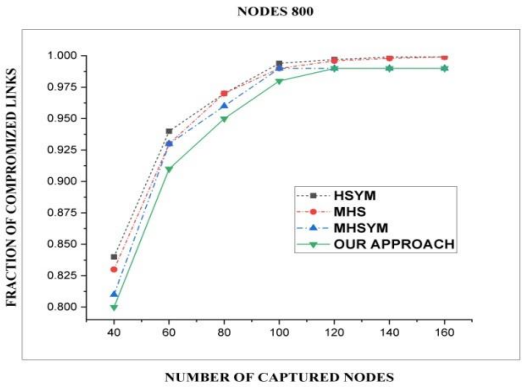


Fig. 4. Simulation Resilience Results of Our Approach Versus HSYM[18], MHS[21], and MHSYM[22] for the 800 Nodes.

TABLE VI. RESILIENCE VALUES FOR 1700 NODES

N	Compromized nodes x	q value	HSYM	MHS	MHSYM	Our Approach
1700	80	37	0.89	0.94	0.87	0.86
	100		0.94	0.9904	0.91	0.9
	120		0.967	0.996	0.95	0.93
	140		0.982	0.998	0.97	0.96
	160		0.9904	0.999	0.99	0.98
	180		0.994	0.999	0.99	0.98
	200		0.997	0.999	0.99	0.98

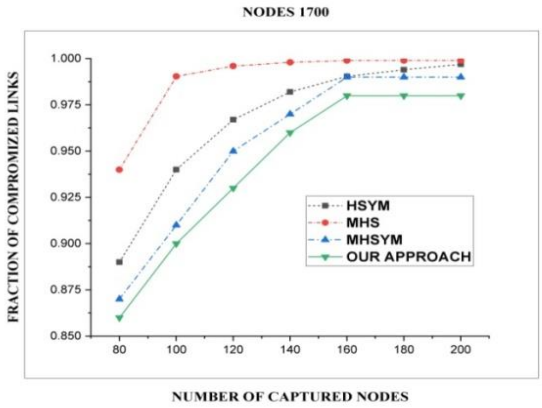


Fig. 5. Resilience Simulationresults of Our Approach Versus HSYM[18], MHS[21], and MHSYM[22] for the 1700 Nodes.

The above graphs and tables prove that our system greatly reduces the probability of compromized network links.Each node receives a different key for its links, and they all also get dynamic keys.

D. Memory Utilization

Here, DGDC is proposed as a centralized key distributor in the proposed system. So, there is no pressure on any network node to maintain all keys in the memory. The IoT node should store only one key that is applied to get the session key from DGDC.

As a result, We can declare that our proposed strategy improves node capture resilience with a combinatorial design.The notations and descriptions of the different parameters used in the article are given in Table VII.

TABLE VII. NOTATIONS OF PARAMETERS

Data related to implementing the Combinatorial designs	Parameter Notation
Blocks (nodes) connected to the IoT network	N
Blocks are generated by SBIBD	B
Remaining Blocks	R
Blocks are generated by HBIBD	H
Number of keys used in each block	k
Key Pool	v
Keys each replicated in the blocks	r
Number of keys intersecting any two blocks	γ

V. DISCUSSION

There is a demand for network security research that is essential due to the upsurge of online transactions. Every user in the transactions believes that the data will be secure and unaltered during transmission. To make secure data and provide reliable keys, a lot of algorithms can be used to provide confidentiality for the data and key-management techniques. In the present work, we are discussing a simple key management algorithm with less time and space complexity compared to the relevant studies on key management algorithms using combinatorial design. We observed that if the network has more than 800 nodes, the

comprised links are reduced when compared to existing techniques. We also mentioned the relevant graphs of resilence in Fig. 3, Fig. 4, and Fig. 5 for various nodes.

VI. Conclusion and Future Work

In this paper, we present a new hybrid combinatorial key distribution scheme for IoT networks that improves the key share probability, scalability, and resilience against capture attacks. In comparison to the other three hybrid methods, our experimental outcomes were better. For all connected nodes, our suggested approach provides the key sharing probability with 1. Every link in the established IoT network can use the same unique key. This paper also provides low resilience values against capture attacks when compared to other schemes. We will also extend this work to reduce the resilience of specific attacks like a man in the middle, Denial of service. We would like to implement it in real networks for better analysis.

References

[1] Alagheband, Mahdi R., and Mohammad Reza Aref. "Dynamic and secure key management model for hierarchical heterogeneous sensor networks." IET Information Security 6.4 (2012): 271-280.

[2] Kouicem, Djamel Eddine, AbdelmadjidBouabdallah, and Hicham Lakhlef. "Internet of things security: A top-down survey." Computer Networks 141 (2018): 199-221.

[3] Stinson, Douglas R., and Scott A. Vanstone. "A combinatorial approach to threshold schemes." SIAM Journal on Discrete Mathematics 1.2 (1988): 230-236.

[4] Stinson, Douglas R. "Combinatorial designs: constructions and analysis." ACM SIGACT News 39.4 (2008): 17-21.

[5] Chien, Hung Yu, Rung-Ching Chen, and Annie Shen. "Efficient key pre-distribution for sensor nodes with strong connectivity and low storage space" 22nd International Conference on Advanced Information Networking and Applications (aina 2008). IEEE, 2008.

[6] Blundo, Carlo, et al. "Perfectly secure key distribution for dynamic conferences" Information and Computation 146.1 (1998): 1-23.

[7] Liu, Donggang, Peng Ning, and Kun Sun. "Efficient self-healing group key distribution with revocation capabilit." Proceedings of the 10th ACM conference on Computer and communications security. 2003.

[8] Chan, Haowen, Adrian Perrig, and Dawn Song. "Random key predistribution schemes for sensor network" 2003 Symposium on Security and Privacy, 2003. IEEE, 2003.

[9] Qian, Sun. "A novel key pre-distribution for wireless sensor networks" Physics Procedia 25 (2012): 2183-2189.

[10] Li, Wei-Shuo, et al. "Threshold behavior of multi-path random key pre-distribution for sparse wireless sensor networks." Mathematical and Computer Modelling 57.11-12 (2013): 2776-2787.

[11] Catakoglu, Onur, and Albert Levi. "Uneven key pre-distribution scheme for multi-phase wireless sensor networks." Information Sciences and Systems 2013. Springer, Cham, 2013. 359-368.

[12] Camtepe S, Yener B “Key distribution mechanisms for wirelesssensor networks: a survey” Rensselaer Polytechnic Institute,Troy,New York, Technical Report,2005, 05-07.

[13] Lee, Jooyoung, and Douglas R. Stinson. "A combinatorial approach to key predistribution for distributed sensor networks" IEEE Wireless Communications and Networking Conference, 2005. Vol. 2.

[14] Ruj, Sushmita, and Bimal Roy. "Key pre-distribution using partially balanced designs in wireless sensor networks" International Journal of High Performance Computing and Networking 7.1 (2011): 19-28.

[15] Ruj, Sushmita, Amiya Nayak, and Ivan Stojmenovic. "Pairwise and triple key distribution in wireless sensor networks with applications" IEEE Transactions on Computers 62.11 (2012): 2224-2237.

[16] Bechkit, Walid, et al. "A highly scalable key pre-distribution scheme for wireless sensor networks" IEEE transactions on wireless communications 12.2 (2013): 948-959.

[17] Bahrami, PoonehNikkhah, et al. "A hierarchical key pre-distribution scheme for fog network." Concurrency and Computation: Practice and Experience 31.22 (2019): e4776.

[18] Camtepe, Seyit A., and BlentYener. "Combinatorial design of key distribution mechanisms for wireless sensor networks" IEEE/ACM Transactions on networking 15.2 (2007): 346-358.

[19] Chakrabarti, Dibyendu, Subhamoy Maitra, and Bimal Roy. "A key pre-distribution scheme for wireless sensor networks: merging blocks in combinatorial design" International Journal of Information Security 5.2 (2006): 105-114.

[20] Kavitha, T., and D. Sridharan. "Hybrid design of scalable key distribution for wireless sensor networks" International Journal of Engineering and Technology 2.2 (2010): 136.

[21] Dargahi, Tooska, Hamid HS Javadi, and Mehdi Hosseinzadeh. "Application-specific hybrid symmetric design of key pre-distribution for wireless sensor networks" Security and Communication Networks 8.8 (2015): 1561-1574.

[22] Akhbarifar, Samira, et al. "Hybrid key pre-distribution scheme based on symmetric design" Iranian Journal of Science and Technology, Transactions A: Science 43.5 (2019): 2399-2406.

Authors’ Profile

G.V.Hindumathi is currently pursuingPh.D. in Jawaharlal Nehru TechnologicalUniversity, Kakinada, India. She is specialized in Internet of Things andNetwork Security. Her research topic is onSecurity issues on Internet of Things.

Dr D.Lalitha Bhaskari works asProfessor in Andhra University,Visakhapatnam, and Andhra Pradesh. Her areas of expertise include: Deep Learning,Network Security, and Image Processing.And she got Young scientist award from byIEI.

Automated Study Plan Generator using Rule-based and Knapsack Problem

Muhammad Amin Mustapa, Lizawati Salahuddin, Ummi Rabaah Hashim
Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka (UTeM)
Durian Tunggal, Melaka, Malaysia

***Abstract*—Undergraduate students are given the flexibility of arranging courses throughout their study duration especially when they are eligible for credit exemption for the courses taken during their diploma study. Issues arise when students arrange their studies manually. Improper course arrangement in the study plan may be resulting some of the selected courses do not correspond to the courses offered, and imbalance credit hours. Hence, this study aims to propose an algorithm to generate an automated and accurate study plan throughout the study duration. A combination of rule-based and knapsack problem were proposed to generate an automated study plan. A quantitative methodology through expert's reviews and questionnaire survey was conducted to evaluate the accuracy of the proposed algorithm. The proposed algorithm shows high accuracy. In conclusion, the combination of rule-based and knapsack problem is appropriate to generate an automated and accurate study plan. The automated study plan generator can help students generate an effective study plan.**

Keywords*—Knapsack problem; rule-based; study plan; undergraduate; credit exemption

I. INTRODUCTION

Study planning is important to ensure the students carry a balance study load in every semester. The balance of courses and the number of credit hours chosen by the students themselves determine the planning of non-burdensome study sessions. Students need to allocate time (also known as student learning hours) for the implementation of all learning activities to achieve the learning outcomes. The student learning hours includes formal meetings (e.g., lectures), guided learning (e.g., tutorials, seminars, internships, and fieldwork), self-directed learning, and preparations for tests and final exams. A balance study load could influence the student's academic performance. Study planning is becoming more critical for undergraduate students who are eligible for credit exemption for the courses taken during their diploma study. When each student has a different number of total credit exemptions, the difference becomes more pronounced. As a result, different study plan is devised for each student. With a well study plan, students may shorten their undergraduate study depending on the total credit exemption.

Study planning is tied to academic rules. For instance, students are allowed to take minimum of 12 credits (exception on the last semester of study), and maximum of 20 credits per semester. Besides, a prerequisite course must be completed prior to another course. Moreover, not all courses are available in every semester. Some courses are only available in odd semesters, while others are only available in even semesters. Therefore, courses should be arranged according to the curriculum structure, and total credit hours per semester should be divided appropriately. The study plan should not interfere with the learning journey.

The field of artificial intelligence (AI) and knowledge-based system has enormous potential for improving simulation modelling support [1, 2]. Due to recent advances in the field of AI, a knowledge-based system has demonstrated its abilities by providing successful solutions in a wide range of applications including in the field of education [2], agriculture [3], manufacturing [4], and health [5]. The system can be used as an alternative to traditional systems, particularly in advisory tasks and symbolic reasoning [6]. It is a subfield of AI that collects data automatically without the assistance of a human expert to solve problems that normally necessitate human intelligence [7, 8].

Rules can be viewed as a simulation of the cognitive behavior of human experts. A rule-based expert system can mimic the ability of human experts to make decisions [9], [10]. They are programmed to solve problems in the same way that humans do, by using stored human information or expertise. Rule-based structures are created to solve specific problems in a given domain. Every domain has its own set of intelligent and reasoning humans that can be modelled and even replaced by automated rule-based systems. A system generator based on a rule engine that uses an improved Rete algorithm was designed to match data objects to perform certain functions through a system generator using rules set by the user (production) [11]. The rule engine's primary responsibility is to match the data objects submitted to the engine with the business rules, activate the business rules based on the current data state, and trigger the operations in the application based on the execution of logic declared in the ruleset. Reference [12] states that the problem of scheduling by minimizing the amount of flow time has attracted more attention from the research community. This is because the lower the total flow time value, the greater the resource utilization and cost savings. In this regard, today's manufacturing environment is quite practical, as it reduces the amount of flow time. Several tasks comprised of some sequences are utilized to determine optimal values for minimizing overall flow time; to provide good solutions as the problem size expands the development of heuristics and meta-heuristics is essential. In the study, a ruled-based heuristic process for determining the sequence with the least total flow

time is proposed. The experimental results show that the proposed approach makes a major contribution to the exceedingly difficult scheduling problem.

The basic principle of all knapsack problem families is to choose a few objects, each with a benefit and weight value, to be packed into one or more capacity knapsacks. Assume there is a group of elements with known weights and values, as well as a pack or bag with a limited capacity for filling the knapsack. A problem known as the knapsack problem is devised to fill the said pack with the elements in such a way that their aggregate sum is possibly the highest without exceeding the pack's ability [13]. Knapsack Problem 0-1 is a popular form of knapsack problem with a wide range of applications, including capital budgeting, project selection, resource allocation, cutting stock, and investment decision-making. As a result, the issue of Knapsack Problem 0-1 optimization has drawn the attention of an increasing number of researchers [14]. GRASP technique was applied to a nurse-scheduling problem where the goal is to optimize a collection of preferred courses to a set of binding constraints [15]. A critical challenge is striking a balance between feasibility and optimality. Construction heuristics, neighborhood search methods, and evolutionary algorithms have all been effectively utilized to solve real scheduling issues. However, there is a frequent conflict between feasibility and solution quality, as well as difficulties in maintaining an appropriate balance between goals. This is solved by employing a knapsack problem, which ensures that the solutions generated by the construction heuristic are simple to fix. A diversification approach and a dynamic assessment criterion improve the optimum combo even further.

Study in [16] developed an automatic course planning system by using ontology and rule-based. The aim was to create a suitable course plan for a group of students according to the course prerequisite requirement, complexity of the course, teaching method, and the duration of the course. However, the course planning system did not include the course scheduling for a complete study duration from year one until end of study duration. Machine learning techniques were used to group students into similar study pattern according to the CGPA achievement and subsequently determine a feasible study path for the forthcoming semester [17]. Specifically, Neural Network algorithm is used for creating CGPA prediction models, and K-means algorithm is applied to group students according to the similarities of their grades in each course. The evaluation of the proposed system revealed that the students have improved their study performance for their ultimate CGPA in graduation. However, the proposed system does not consider the duration of the study completion, the course prerequisite requirement, and total number of credits in a semester. Moreover, [18] in their research work addressed the issue of determining the ideal set of courses to provide students with in a particular semester, while taking into account the required courses and the availability of teachers to teach those courses. The use of *CourseScheduler*, *IApplet* and *AdmValidatorApplet* function altogether helps the authors to achieve their aim successfully. However, the research focused on the creating a schedule of classes that aid the department administrative in the course scheduling rather than the study plan for students. The method to assist students generating study plan is lacking. Based on these limitations, this study aims to propose and validate an algorithm for compiling study plans throughout the study duration. A more in-depth investigation was conducted to assess the method's accuracy and usefulness.

II. BACKGROUND

A direct entry student is defined as a student who pursues a degree from a particular institution or a university with a particular completed diploma degree. Compared to direct admission students, direct entry students are allowed to make credit exemption. Credit exemption is a provision of the academic regulations under the semester system that aims to facilitate student mobility. For an instance, students must complete a diploma with at least a 3.00 CGPA from an institution and the courses pursued must be recognized by the senate as equivalent and meet the curriculum requirements of the program pursued or in their respective field of study. Credit exemption may be granted to students who have taken equivalent courses and passed with a minimum grade of C using the university's grading system, provided that at least 80% of the learning content is equivalent. The amount of credit exemption allowed should not exceed 30% of the total credits of the graduating requirements.

Academic handbooks have become a reference for students, containing important information about students' curriculum structure according to a specific program. Courses are divided into four categories, namely general module (W), core module (P), specialization module (K), and free module (E). Table I explains the course category. All courses are categorized as W and are not allowed for credit exemption.

TABLE I. COURSE CATEGORY

Component	Code	Meaning	Credits
General Module	W	University Compulsory Courses, which are a group of important Courses determined by the Senate and made compulsories for all students.	14
Program Core Module	P	Mandatory courses to meet the requirement of Bachelor of Computer Science.	45
Final Year Project			6
Industrial Training			12
Specialization Module	K	Specialization courses to a specific major of an academic program.	30
Free Module	E	Elective Courses that are offered to deepen an academic program.	13
Total Credits			120

TABLE II. COURSE CODE AND NAME

Code	Course Code	Course
P1	BITU 2913	Workshop I
P2	BITU 3973	Final Year Project I
P3	BITU 3983	Final Year Project II
P4	BITP 1113	Programming Technique
P5	BITI 1113	Artificial Intelligence
P6	BITS 1313	Data Communication and Networking
P7	BITP 3113	Object Oriented Programming
P8	BITP 2213	Software Engineering
K1	BITU 3923	Workshop II
K2	BITI 2213	Knowledge Based System
K3	BITI 3413	Natural Language Processing
K4	BITI 2223	Machine Learning
W1	BLHW 1442	English for Academic Purposes
W2	BLHW 2452	Academic Writing
W3	BLHW 3462	English for Professional Interaction
W4	BKK ---1	Co-Curriculum I
W5	BKK ---1	Co-Curriculum II
E1	BLHC 4302	Critical and Creative Thinking

Table II shows the list of course code and name. Students must meet all components of the code to complete a total of 120 credit hours.

III. ALGORITHM IMPLEMENTATION INTO UNIVERSITY RULE

Rule-based expert system is based on knowledge that collects a range of factual information, and makes actions through interpretation from a set of predefined rules [19]. Certain courses have rules that must be followed to complete the semester. According to the rules, the proposed algorithm will decide whether or not to include the course. For the arrangement of study plans, a new rule is proposed to control the arrangement of schedules according to the selected semester. The selection of courses is based on the current semester offers and availability through the knapsack problem method until the credit hour rate reaches a predetermined limit.

Normal students must complete all 120 credits in a minimum of 7 semesters. However, direct entry students have the option to shorten the semester depending on the total number of credit exemption approved. Alternatively, students can stay with 7 semesters as offered with lower credit hours. The first step is to determine the maximum number of semesters according to the total number of credit exemption approved. Then, the total credit hour in a semester is calculated to ensure a balance credit taken by students in every semester, and to ensure that the total credit hours to be taken do not exceed the stipulated conditions. Fig. 1 depicts the semester calculation step as well as the total credit hours in a semester according to total credit exemption approved.

Equation (1) shows the calculation of total credit hours in a semester (*tch*).

$$tch = \frac{120-\text{tce}-12}{(\text{ts}-1)} \quad (1)$$

where 120 is the minimum graduating credit, *tce* is total credit exemption, 12 is the Industrial Training credit, and *ts* is the total number of semesters.

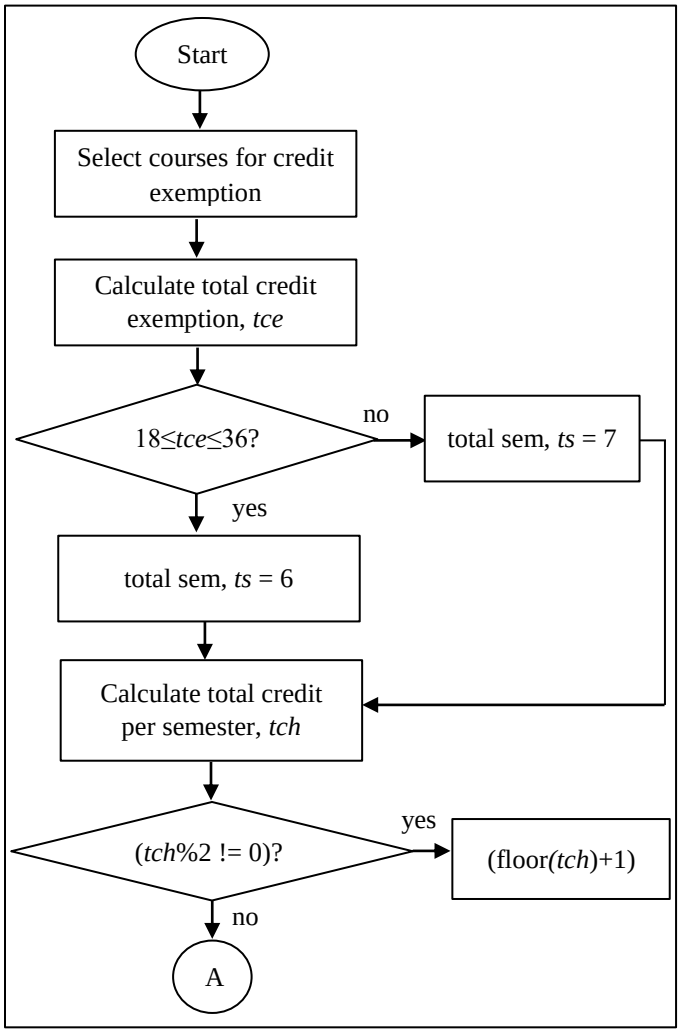


Fig. 1. Total Semester and Credit Hour according to Total Credit Exemption.

Industrial Training is a mandatory requirement for students at the end of the semester before graduation. Hence, the industrial training credit and the semester are deducted to calculate *tch*. To make the calculation method simpler, the maximum amount generated will be added to a value of 1 for the semester when the calculation result produces a decimal number and the decimal part is removed as shown in Fig. 1. This is because the number of available credits offered varies and there are no credit hours in decimal form. Using this formula, the total credit hours will not fall below the semester's minimum total credit of 9 and will not exceed the semester's maximum total credit of 20.

Credit exemption is permitted for program core courses. Not all courses can be exempted, including those courses with W category. Workshop I (P1), Workshop II (K1), Final Year Project I (P2), and Final Year Project II (P3) are project-based courses that cannot be exempted. P1 and P2 is the prerequisite course of K1 and P3, respectively. Moreover, K1 is the prerequisite course of P2. As they are offered once a semester and have pre-requisites, these courses should not be taken lightly. The proposed algorithm has set some rules based on the number of semesters. Each rule has unique characteristics

for each course. The course is thus removed from the list of available courses because it has become a rule that must be followed. Using the proposed rule-based approach, several courses must be prioritized to ensure the planned flow runs smoothly. Overall, rule-based algorithm is applied to:

- determine number of semesters study
- prioritize University Compulsory Courses to be arranged in the semester according to the program curriculum structure.
- prioritize program core courses according to the pre-requisite and semester offered.
- prioritize specialization courses according to the pre-requisite and semester offered
- prioritize English courses according to the pre-requisite and semester offered.

The parameters used for configuring the rule-based algorithm include the total credit hours, exempted courses and their credit hours, course prerequisites, program curriculum structure and course details. The course details including course name, course code, course category, credit hours, and the semester offered according to odd or even semesters.

A. General Rule for Seven Semesters of Study

Students who only receive credit exemption ranging from 3 to 15 credit hours are advised to complete seven semesters of study. This is because the number of exemption hours is insufficient to reduce the study time. However, students can reduce the credit hours for the coming semesters. The rule prioritizes the courses categorized as W and K to be arranged in the semester according to the curriculum structure as depicted in Fig. 2.

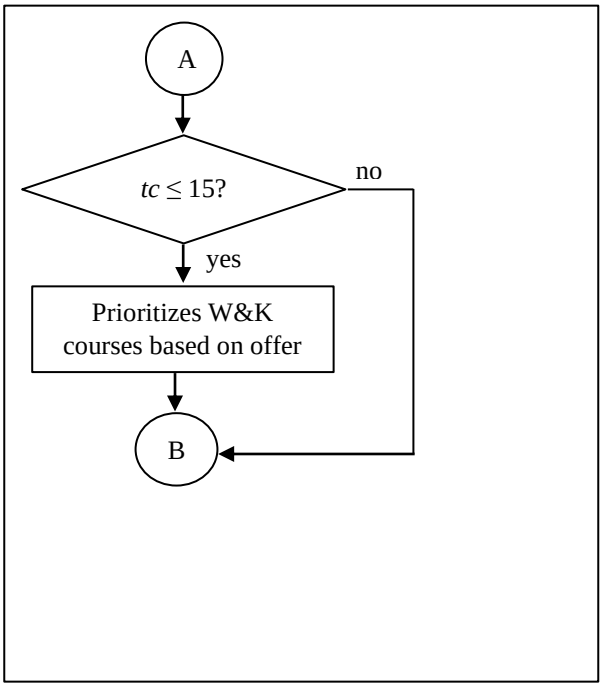


Fig. 2. 7-Semester Rules.

B. General Rule for Six Semesters of Study

Fig. 3 illustrates how the rules for 6 semesters are applied. For all programs offered at the faculty, workshops (P1 and K1) are the main course at the core of the program. The P1 course is available in both semesters, but the K1 course is only offered in the odd semester. Therefore, students are encouraged to take P1 early in the semester so that they can enroll K1 in the subsequent odd semester with 3rd year students. Then, students are allowed to take final project courses in the following semester. Moreover, P4 (Programming Technique) is the pre-requisite course of P1. In this way, planning to shorten the semester is more structured because the core courses can be enrolled in the appropriate semester.

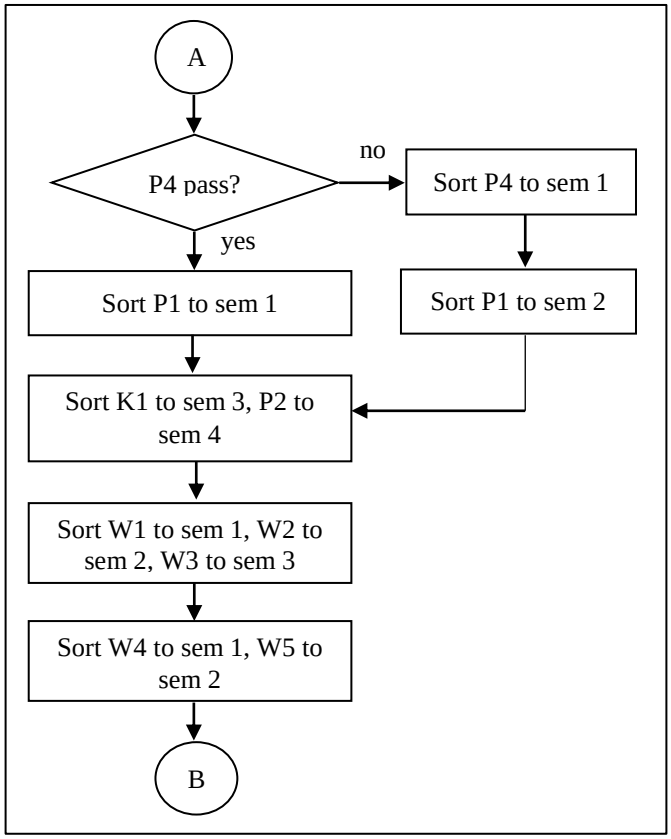


Fig. 3. 6-Semester Rules.

P1 and K1 are given priority because they are the backbone of the study plan and are divided into six semesters. According to the curriculum structure, these workshop courses are only offered in odd semesters. In semester 1, the algorithm will check the pre-condition status of the P1 course first before allocating the course in a semester. If the requirements are not complied with, students will first consider the P4 and change P1 to the second semester. The group for K1 comprises direct entry students and normal entry students. Hence, the course must be offered in semester 3 to ensure the direct entry students can be assigned in groups. This is equivalent to semester 5 of normal students. Next, English courses are placed in the earlier semester such as English for Academic Purpose (W1), sorted to semester 1, English for Academic Purposes (W2), sorted to semester 2, and English for

Professional Interaction (W3) at semester 3. The next rule ensures the selection of co-curriculum courses. The method is the same as the prerequisites by ensuring that co-curriculum courses are not taken in the same semester and Co-Curriculum II (W5) does not precede Co-Curriculum I (W4).

C. Course Specialization Rules

In addition, specialization based on the program taken by the students is emphasized. This is because these courses are only concentrated among the same programs. The students are not permitted to join specialization classes of other programs. Since specialization courses are offered at a particular semester, a rule is made to allow and ensure specialization courses are taken during the semester where the courses are offered. This ensures the students are following the correct guidelines throughout their study. The rules have limited the students to take 3 or 4 courses per semester to ensure that their study schedule is bearable during the semester. Therefore, maximum specialization courses are set based on the proposed rules. There are several additions to the rules for certain programs such as Bachelor of Computer Science (Database Management) with honors (BITD), Bachelor of Computer Science (Computer Networking) with honors (BITC), and Bachelor of Computer Science (Artificial Intelligence) with honors (BITI). For instance, for the BITI program, the P5 (Artificial Intelligence) course is a prerequisite that must be met. The P5 course affects other courses like K2 (Knowledge-Based System), K3 (Natural Language Processing), and K4 (Machine Learning). Fig. 4 illustrates the additional rules of BITI program.

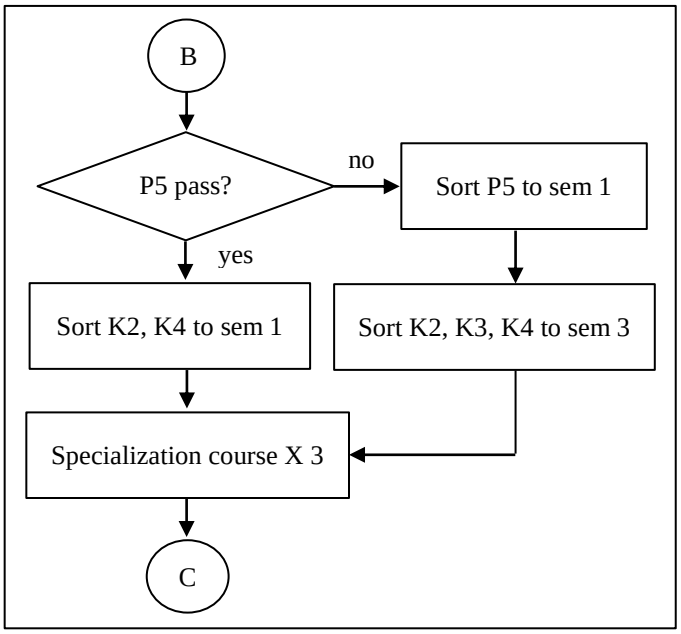


Fig. 4. BITI Specialiazation Rules.

D. Course Availability Rule

The availability of a course should be considered before carrying out this proposed rule. This is because specialization courses need to be sorted accordingly. It is important to offer the courses according to odd or even semesters so that students are not left behind when the courses are offered. If the courses are only offered in odd semesters, they will not be available in even semesters, and vice versa. Some courses are open to other programs in other semesters. Students can plan ahead of time to enter the classes indirectly. This arrangement is based on the lean and the year of the offers to correspond to the students' year of study. This arrangement must be made to ensure that students take a diverse range of courses while also meeting the required credit hours. The proposed algorithm will ensure the availability of a course's semester whether it is in an even or odd semester only or both.

Fig. 5 shows a continuation of the previous compilation of rules. The next rule stipulates that the specialization courses should be included in a particular semester. This is because the semester arrangement is short, and some courses need to be taken first. Specialization courses according to a particular program are usually not offered in other programs; hence should be prioritized in the compilation.

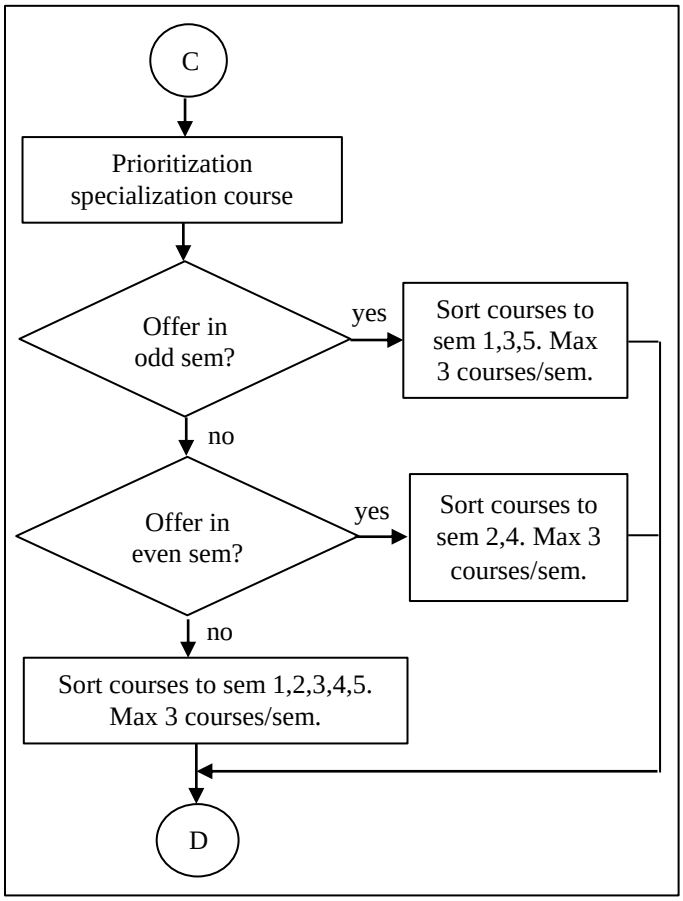


Fig. 5. Specialization Course Flow.

E. Sorting Courses using Knapsack Problem 0-1

The Knapsack Problem 0-1 is applied to fill the remaining number of credit hours from the rule-based algorithm until the total credit hour limit is reached. Equation (2) shows the equation for the Knapsack Problem 0-1.

$$max \sum_{i=1}^{n} xi * pi$$

$$\sum wi\ xi\ \leq c$$

$$xi\ \in \{0,1\}, i = 1, \ldots\ldots, n.$$

$$pi > 0, wi > 0, c > \qquad (2)$$

where i represents course (xi = 1 for selected course, whereas xi = 0 for unselected course), n is a number of total courses, wi is weight, pi is profit which is the credit hour of a particular course, and c is the required remaining credit hours to fulfil the total credit hours per semester. The algorithm will select the highest and most appropriate credit hours that can be adjusted for the number of credit hours remaining. The election results made by the proposed algorithm are entered into the semester. This process is repeated until the total number of credit hours reaches a maximum. This process continues to compile for the next semester. Fig. 6 depicts the flow of the knapsack problem where the process is repeated until the number of hours and courses for each semester reaches the maximum rate.

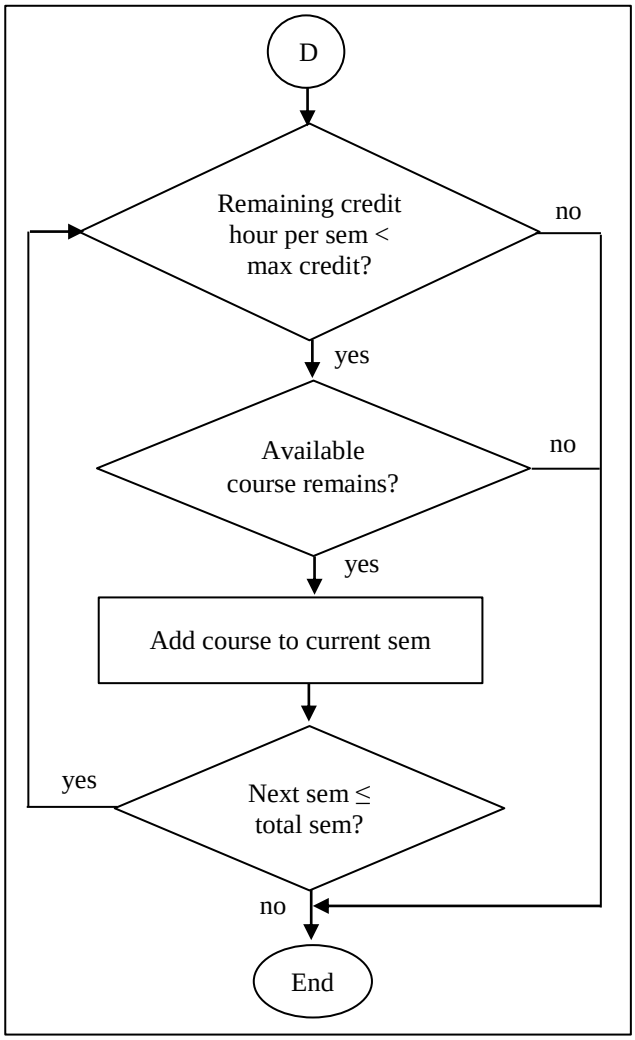


Fig. 6. Sorting Courses using Knapsack Problem 0-1.

IV. Automated Study Plan Generator Prototype

The creation of the courses details for a complete curriculum structure is the first stage in constructing this prototype. The course details include course code, course name, prerequisites course, credit hours, and semester of offering are the required input as shown in Fig. 7.

Fig. 7. Course Details.

The total credit hours exempted is then determined as shown in Fig. 8. This stage is crucial since it serves as the prototype's major support structure. The total credit hours exempted must not exceed the maximum credits set by the university. The prototype shows warning when total credit hours exempted exceed the maximum credits to prevent students from making mistakes.

Fig. 8. Exempted Courses and Credit Hours.

Subsequently, the algorithm will determine the number of semesters and arrange the courses that are appropriate for the student. According to the parameters given by the algorithm, new courses will be substituted for the exempted courses. Fig. 9 shows the example of courses plan generated from the automated study plan generator.

Fig. 9. Example of Courses Plan.

V. METHODS

A. Expert Reviews

Expert review was conducted to evaluate the suitability and accuracy of the proposed algorithm. For this study, experts consist of lecturers who had experience as academic advisors in a faculty. In total, four experts representing various academic program were participated in the review. The experts were contacted in advance to obtain information about their experience as academic advisors and to obtain their consent to become experts. Each expert was chosen from different departments to ensure that the rules established for each program were followed correctly. Then, the test case was sent via email.

The preparation of test cases was planned following the program to be given to experienced academic advisors. Test cases were organized based on the study plan generated from the proposed algorithm. To ensure accuracy, respondents were allowed to test the automated study plan generator prototype at random. The test cases were divided into 3 sections. Section A contains five test cases of total credit exemption between 3 and 15 credit hours which allows students to take 7 semesters of study. Section B contains five test cases of total credit exemption between 18 and 36 credit hours which allows students to take 6 semesters of study. Section C contains 3 test cases based on the random credit exemption course selected by the respondents. The total credit exemption between 3 and 36 credit hours. In total, 32 test cases were distributed to the experts.

B. Testing

Testing was conducted to identify bugs in the proposed algorithm. Testing helps in understanding and refining the given requirements [20]. It is the practice of comparing a piece of software's behavior to the predetermined and expected behavior established during the development phase. This method of testing accuracy was accomplished through the use of a study plan generated by the prototype. This test was run for each program several times to identify any problems that may have arisen. This test was performed independently to ensure that the study plans produced met the study's objectives.

C. Questionnaire Survey

A user acceptance survey was developed with Google Forms and sent through messages to respondents. The survey was distributed to direct entry students who are aware of the direct entry concept and procedures. The questionnaire items were separated into sub-categories to acquire a clear understanding and accountability of evaluations and comments at the next step. The technology acceptance model (TAM) created by Davis was used in this study to evaluate the behavior of persons by using one generally known theory on the actual use behavior of utilizing new technology [21]. The influences on the intention of using the prototype were based on the individual's perceived ease of use (EU). The capability of the prototype (CP) was determined in terms of features and results generated to leverage user needs for study plan activities. Attitude (ATT) was thought to assist in meeting the needs of the users and hence influenced the attitude created toward the prototype. The perceived usefulness (PU) ensured that the developed prototype received a response in terms of use and usage behavior. Lastly, the student intends to use (IU) was created to determine the extent to which prototype requirements were developed to address existing problems. Fig. 10 illustrates the revised TAM model that specifically explains the computer acceptance determinants that are general and capable of explaining user behavior toward the automated study plan generator.

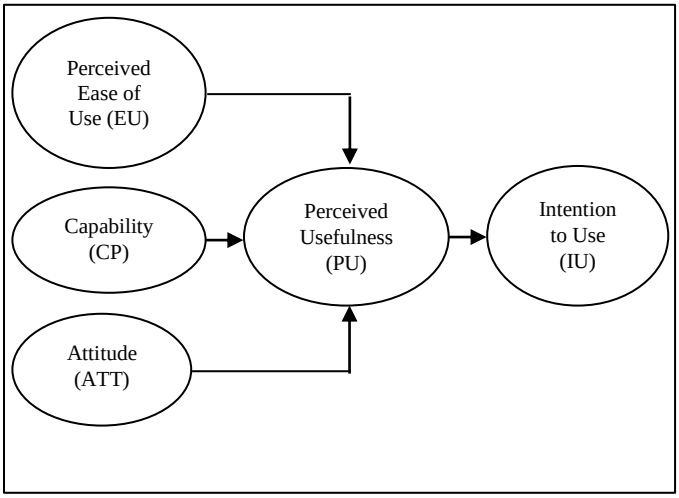


Fig. 10. Revised TAM for Automated Study Plan Generator.

The questionnaire used a five-point Likert scale of 1 to 5, 1 refers to strongly disagree and 5 refers to strongly agree. The questionnaire consists of 19 items. Respondents were instructed to use the automated study plan generator first to provide an overview of the prototype. Next, the respondents were required to answer the questionnaire survey. Each aspect presented was analyzed to gain the respondents' acceptance of the prototype to achieve the objectives.

D. Data Analysis

Test case results from experts and testing were analyzed using a confusion matrix to evaluate the accuracy. Table III shows the aspects to calculate the accuracy of the matrix by taking the average values across the "main diagonal".

The formula to calculate the accuracy based on the confusion matrix is shown in (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

TABLE III. CONFUSION MATRIX FOR BINARY CLASSIFICATION

		True Class	
		Positive	**Negative**
Predicted Class	**Positive**	**TP**: Expected outcome was YES, and the actual outcome was also YES.	**FP:** The expected outcome was YES, and the actual outcome was NO.
	Negative	**FN:** The expected outcome was NO, and the actual outcome was YES.	**TN**: The expected outcome was NO, and the actual outcome was also NO.

On the other hand, a descriptive analysis (Mean ± SD) and correlation analysis were performed to analyze the data collected form the questionnaire survey.

VI. RESULTS

Table IV shows the results from the expert review. In total, four experts representing various program were participated in the review. Each expert evaluated thirteen cases of a particular program. The results indicate that the automated study plan generator generates highly accurate study plan between 0.99 to 1 accuracy.

According to an expert who evaluated BITI program, the P1 (Workshop 1) course should be scheduled in the second semester so that students can learn how to tackle their studies at UTeM first. If a student successfully exempted for more than 18 credit hours, the algorithm rules place P1 course in Semester 1. They can be changed because the P1 course is available in both, and a new course will be taken. As for the BITC program, the arrangement of courses developed in the automated study plan generator is good, except that there is a problem with the K1 (Workshop II) course that is supposed to be taken before the end of Year 2. This is because the outcome of the arrangement produced depends on the P6 (Data Communication and Networking) course whether it is excluded or not. This plays an important role in compiling the study plan, but it is not stated in the handbook. Human error occurs where the availability of elective courses is incorrectly set causing the accuracy of the program to decrease.

Overall, the study plan generated from the proposed algorithm has high accuracy. It is very useful for new direct entry students to obtain an initial overview of the preparation of study plans at the beginning of the semester. The courses offered also depend on the quota set by the faculty, which forces students to change their study plans in the event of a change.

A. Testing

Table V shows 28 manual testing results from various programs. Random course selection reveals that the automated study plan generator prototype has an accuracy of 0.999 on an average. The accuracy of the manually tested program has given a value of 1 except for the BITE program. This is because when the BITE program is shortened; students must merge three even semesters into two semesters. This causes the generated study plan exceeds the total credit hours. If a student is exempted from 18 credit hours, but the courses provided in the second semester are not reduced, the generated study plan will be unbalanced credit hours.

B. Questionnaire Survey

Forty-four direct entry students have participated in the survey. These students were from Semesters 2 and 6. The female and male respondents were 43.2% and 56.8% respectively. BITS program had the highest percentage of 45.5%, followed by BITI and BITD at 15.9%. Besides, there are 11.4% students from the BITC program and 9.1% from the BITM program. Lastly, there are 2.3% of students from the BITZ program.

TABLE IV. RESULTS FROM EXPERT REVIEW

Program	TP	TN	FP	FN	Total	Accuracy
BITI	539	0	7	0	546	0.99
BITM	546	0	0	0	546	1.00
BITC	538	0	8	0	546	0.99
BITS	544	0	2	0	546	0.99

TABLE V. TESTING RESULTS

Program	TP	TN	FP	FN	Total	Accuracy
BITI	252	0	0	0	252	1
BITS	126	0	0	0	126	1
BITM	210	0	0	0	210	1
BITC	210	0	0	0	210	1
BITZ	168	0	0	0	168	1
BITE	125	0	1	0	126	0.992
BITD	84	0	0	0	84	1
Total	**1175**	**0**	**1**	**0**	**1176**	**0.999**

TABLE VI. DESCRIPTIVE ANALYSIS OF USER ACCEPTANCE CONSTRUCTS

Construct	Mean ± SD
Perceived ease of use (EU)	4.301 ± 0.610
Perceived usefulness (PU)	4.291 ± 0.624
Capability (CP)	4.369 ± 0.561
Attitude (ATT)	4.348 ± 0.618
Intention to use (IU)	4.242 ± 0.619

Table VI shows a descriptive analysis of the acceptance test constructs. All mean values are greater than 4.2, indicating that respondents have a generally positive opinion of the automated study plan generator. A total of 96% of respondents agreed that the automated study plan generator is capable of producing a study plan that meets the specified requirements. The majority of respondents (92%) rated all items under attitude and perceived ease of use constructs on a scale of 4 (agree) to 5 (strongly agree). All respondents also agreed on the automated study plan generator's perceived usefulness. Lastly, the automated study plan generator would be used by more than 90% of the respondents.

Correlation analysis of the acceptance test between constructs is shown in Table VII. The results indicate all the constructs show a positive and strong correlation (exceeding 0.5), with all correlations significant at the $p<0.01$ level. The relationship between capability (CP) and intention of use (IU) is 0.862, indicating that the two are highly correlated. The finding implies that user intention is based on the capabilities of the prototype to assist users in achieving the goal of use.

TABLE VII. CORRELATION ANALYSIS OF USER ACCEPTANCE CONSTRUCTS

Construct	EU	PU	CP	ATT	IU
EU	1				
PU	0.842	1			
CP	0.671	0.743	1		
ATT	0.638	0.774	0.769	1	
IU	0.600	0.753	0.862	0.801	1

Following that is perceived ease of use (EU) concerning perceived use (PU), with a high correlation between the two constructs, demonstrating that the prototype is simple to understand and provides convenience to the user.

VII. Discussion

An algorithm for compiling study plans was proposed and validated in this study. Rule-based and knapsack problem were applied in compiling student learning plans. The rule-based method is utilized to optimize the courses that students must take during the semester as specified by the faculty. These courses have been planned based on the total number of credit hours exempted. There are crucial courses that must be prioritized based on the semester to guarantee that students do not miss out and create a change in the intended number of semesters. Besides, the knapsack problem used in this study is intended to select courses that are not included in the rules and can be put into a table based on credit hours and the desired offer. The courses to be chosen are balanced according to the number of hours allotted. As a result, a study plan that satisfies the prerequisites is created. These two approaches are ideal for dealing with this issue. This is because significant courses can be certain of their offer, while other courses are offered following the correct offer. The planned structure qualifies for making a study plan. The results from expert reviews and testing reveal that the automated study plan generator prototype has an accuracy of 0.999 on an average. Moreover, most of the respondents participated in the user acceptance survey have a generally positive opinion of the automated study plan generator in term of ease of use, usefulness and capability. The automated study plan generator would be used by more than 90% of the respondents.

The results produced from this study could provide valuable contributions to the undergraduate students to plan their course schedule prior to their graduation. The process of organizing learning can be more effectively implemented using the proposed algorithm. Students will not be overburdened and will be able to increase the consistency of their learning output in the coming semester by finding suitable learning arrangements. It has the potential to indirectly improve student learning performance.

This study has certain limits and problems. If the rule is incorrect or not written in the academic handbook, it can disrupt the schedule's arrangement. This is because the rules cannot be followed, resulting in a wild and incorrect arrangement. It will stymie students' planning and make new arrangements difficult. When constructing the study plan, Knapsack Problem 0-1 acts greedily to avoid this problem by using rules to ensure the algorithm obeys the established limitations. Knapsack Problem 0-1 will continue to produce results based on the number of credit hours without following the course codes. Moreover, human error is unavoidable when conducting studies, which reduces the accuracy of the results. The combination ruled-base and knapsack problem algorithm assures that the study schedule can be organized properly. Prerequisites can be met, credit hours can be allocated in a balanced manner, and courses can be arranged according to the offers by the faculty, all in one system. With the study's findings, any desired method can be constructed in the future. To solve this problem, rule-based and knapsack problem are appropriate. This is because each course must follow all rules, and voids can be filled by the knapsack problem with greedily picked courses to fulfil the prerequisite credit hours. It also relies on the availability of courses offered in the semester. The results have a high level of accuracy and can be used by academic advisors and new direct entry students to arrange their schedules.

VIII. Conclusion

This study had successfully proposed and validated an algorithm for compiling study plans by using rule-based and knapsack problem. Based on the results and analysis, it is possible to conclude that the accuracy of the algorithm based on the rule-based and knapsack problem to generate study plan is high. Survey respondents believe that the proposed algorithm can assist them in creating and designing study plans. The majority of respondents are interested in using the automated study plan generator. Finally, it can be seen that both students and academic advisors can benefit from the automated study plan generator to arrange their study plans.

Acknowledgment

This study is supported by the Universiti Teknikal Malaysia Melaka research grant (JURNAL/2020/FTMK/Q00055).

References

[1] T. R. Hill and S. D. Roberts, "A prototype knowledge-based simulation support system," Simulation, vol. 48, no. 4, pp. 152–161, 1987, doi: 10.1177/003754978704800407.

[2] H. Yang, M. Anbarasan, and T. Vadivel, "Knowledge-Based Recommender System Using Artificial Intelligence for Smart Education," J. Interconnect. Networks, vol. 2143031, 2022.

[3] M. Á. Rodríguez-García, F. García-Sánchez, and R. Valencia-García, "Knowledge-Based System for Crop Pests and Diseases Recognition," Electronics, vol. 10, no. 8, p. 905, 2021.

[4] M. R. Khosravani, S. Nasiri, and T. Reinicke, "Intelligent knowledge-based system to improve injection molding process," J. Ind. Inf. Integr., vol. 25, no. August 2021, p. 100275, 2022, doi: 10.1016/j.jii.2021.100275.

[5] S. Bashir, A. A. Almazroi, S. Ashfaq, A. A. Almazroi, and F. H. Khan, "A Knowledge-Based Clinical Decision Support System Utilizing an Intelligent Ensemble Voting Scheme for Improved Cardiovascular Disease Prediction," IEEE Access, vol. 9, pp. 130805–130822, 2021, doi: 10.1109/ACCESS.2021.3110604.

[6] W. Y. Zhang, S. B. Tor, and G. A. Britton, "A prototype knowledge-based system for conceptual synthesis of the design process," Int. J. Adv. Manuf. Technol., vol. 17, no. 8, pp. 549–557, 2001, doi: 10.1007/s001700170137.

[7] F. Mustapha, N. Ismail, S. M. Sapuan, Z. Noh, and A. Samsuri, "Development of a prototype knowledge-based system for troubleshooting of aircraft engine and parts - A case study of Cessna Caravan," Int. J. Mech. Mater. Eng., vol. 5, no. 1, pp. 36–42, 2010.

[8] I. H. Sarker, "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems," SN Comput. Sci., vol. 3, no. 2, pp. 1–20, 2022, doi: 10.1007/s42979-022-01043-x.

[9] G. Engin et al., "Rule-based expert systems for supporting university students," Procedia Comput. Sci., vol. 31, pp. 22–31, 2014, doi: 10.1016/j.procs.2014.05.241.

[10] I. H. Sarker, A. Colman, J. Han, and P. Watters, "Context-Aware Rule-Based Expert System Modeling BT - Context-Aware Machine Learning and Mobile Data Analytics: Automated Rule-based Services with Intelligent Decision-Making," I. Sarker, A. Colman, J. Han, and P.

Watters, Eds. Cham: Springer International Publishing, 2021, pp. 129–136.

[11] K. Qu, T. Gong, and J. Shao, "Design and implementation of system generator based on rule engine," Procedia Comput. Sci., vol. 166, pp. 517–522, 2020, doi: 10.1016/j.procs.2020.02.054.

[12] S. S. Raghavan, "Rule Based Heuristic Approach for Minimizing Total Flow Time in Permutation Flow Shop Scheduling," Teh. Vjesn., vol. 22, no. 1, pp. 25–32, 2015, doi: 10.17559/TV-20130704132725.

[13] D. Sapra, R. Sharma, and A. P. Agarwal, "Comparative study of metaheuristic algorithms using Knapsack Problem," Proc. 7th Int. Conf. Conflu. 2017 Cloud Comput. Data Sci. Eng., pp. 134–137, 2017, doi: 10.1109/CONFLUENCE.2017.7943137.

[14] J. Lv, X. Wang, M. Huang, H. Cheng, and F. Li, "Solving 0-1 knapsack problem by greedy degree and expectation efficiency," Appl. Soft Comput. J., vol. 41, pp. 94–103, 2016, doi: 10.1016/j.asoc.2015.11.045.

[15] M. D. Goodman, K. A. Dowsland, and J. M. Thompson, "A grasp-knapsack hybrid for a nurse-scheduling problem," J. Heuristics, vol. 15, no. 4, pp. 351–379, 2009.

[16] R. O. K. Base and P. Nilaphruek, "Automatic Course Planning System Using Rule-Based Ontological Knowledge Base," Int. J. Comput. Internet Manag. Vol.23, vol. 23, no. 1, pp. 16–23, 2015.

[17] N. Chanamarn and K. Tamee, "Enhancing Efficient Study Plan for Student with Machine Learning Techniques," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 3, pp. 1–9, 2017, doi: 10.5815/ijmecs.2017.03.01.

[18] T. Sobh, S. Patel, and R. Cousen, "Course Scheduler : An Automated Schedule Generator," 2007.

[19] X. Wang, Y. Bai, C. Cai, and X. Yan, "A production rule-based knowledge system for software quality evaluation," ICCET 2010 - 2010 Int. Conf. Comput. Eng. Technol. Proc., vol. 6, pp. 208–211, 2010, doi: 10.1109/ICCET.2010.5486303.

[20] C. Klammer and R. Ramler, "A Journey from Manual Testing to Automated Test Generation in an Industry Project," Proc. - 2017 IEEE Int. Conf. Softw. Qual. Reliab. Secur. Companion, QRS-C 2017, pp. 591–592, 2017, doi: 10.1109/QRS-C.2017.108.

[21] R. Rauniar, G. Rawski, J. Yang, and B. Johnson, "Technology acceptance model (TAM) and social media usage: An empirical study on Facebook," J. Enterp. Inf. Manag., vol. 27, no. 1, pp. 6–30, 2014, doi: 10.1108/JEIM-04-2012-0011.

Combining Multiple Classifiers using Ensemble Method for Anomaly Detection in Blockchain Networks: A Comprehensive Review

Sabri Hisham, Mokhairi Makhtar and Azwa Abdul Aziz
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, 22000 Terengganu, Malaysia

***Abstract*—Blockchain is one of the most anticipated technology revolutions, with immense promise in various applications. It is a distributed and encrypted database that can address a range of challenges connected to online security and trust. While many people identify Blockchain with cryptocurrencies such as Bitcoin, it has a wide range of applications in supply chain management, health, Internet of Things (IoT), education, identity theft prevention, logistics, and the execution of digital smart contracts. Although Blockchain Technology (BT) has numerous advantages for Decentralized Applications (DApps), it is nevertheless vulnerable to abuse, smart contract failures, security, theft, trespassing, and other concerns. As a result, using Machine Learning (ML) models to detect anomalies is an excellent way to detect and safeguard blockchain networks from criminal activity. Adapting ensemble learning methods in ML to create better prediction outcomes is a viable approach for anomaly identification. Ensemble learning, as the name implies, refers to creating a stronger and more accurate classification by combining the prediction results of numerous weak models. As a result, an in-depth evaluation of ensemble learning methodologies for anomaly detection in the blockchain network ecosystem is applied in this paper. It comprises numerous ensemble methods (e.g., averaging, voting, stacking, boosting, bagging). The review collects data from three established databases, which are Scopus, Web of Science (WoS), and Google Scholar. Specific keywords are employed, such as Blockchain, Ethereum, Bitcoin, Anomaly Detection, and Ensemble Learning, employing advanced searching algorithms. The results of the search found 60 primary articles from 2017 to 2022 (30 from Scopus, 20 from the WoS, and 10 from Google Scholar). Based on these findings, we decided to divide our debate into three primary themes: (1) the fundamentals of Blockchain Technology (BT), (2) the overview of ensemble learning, and (3) the integration and analysis of ensemble learning in blockchain networks for anomaly detection. In terms of awareness and knowledge, the results are also discussed in terms of what they mean and where future research should go.**

Keywords*—Blockchain; Ethereum; Bitcoin; ensemble; anomaly detection

I. INTRODUCTION

Nowadays, most agencies have started evaluating Blockchain Technology (BT) in various sectors such as pharmaceuticals, automotive, agri-food, livestock, supply chain, health, and government digital initiatives [1]. This scenario has an impact in the context of traceability, transparency, and trustworthiness values in distributed and decentralized ecosystem environments [2]. A Blockchain operates based on a data structure storage method consisting of blocks that are interconnected with each other using a cryptography hash mechanism. Technically, each block stores information such as timestamp, Merkle root, nonce, previous hash and difficulty in the block header [3]. From the point of view of decentralized Blockchain applications, the world of cryptocurrency has become popular and dominant. Thus, Bitcoin BT has forged success by producing the first cryptocurrency application. It is different from Ethereum, which introduced smart contracts, and Ether has been declared the second largest cryptocurrency after Bitcoin [4]. Additionally, Ethereum was created to address the Bitcoin protocol's functional insufficiency [5]. Technically, the Ethereum network hosts smart contracts, which are collections of code that run on the Blockchain and carry out a set of instructions. These contracts are what power Decentralized Applications (DApps), which are akin to smartphone apps that operate on Google (Android) or Apple (iOS) operating systems.

In a public blockchain network, all transactions are transparent and are publicly available. Hence, anyone in the network can examine these transactions and may cross-verify any fraudulent behavior. Along with its rapid development, BT has encountered several security issues and shortcomings, including majority attacks, forking, and bugs in smart contracts. Wallet attacks, Ponzi Schemes, Proof of Work (PoW) vulnerabilities, and crypto-jacking are all challenges that need to be addressed. For instance, the Ethereum Blockchain has increased in prominence. Nevertheless, it has been beset by security vulnerabilities such as phishing scam, which has accounted for nearly half of all criminality on the platform since 2017 [6]. Therefore, for an efficient functioning of a blockchain network, it is vital to detect these vulnerabilities in the most precise and timely manner. To enable the successful identification and prediction of such attacks over Blockchain, the field of anomaly detection models in the Machine Learning (ML) method for Blockchain comes into play.

In general, an attempt to detect an anomaly in a pattern or thing that is different from the norm is termed anomaly detection. [7]. This demonstrates that combining ML and BT has a good impact and is widely employed in industries such as automotive, health, decentralized finance (DeFi), supply chain, agriculture, and the Internet of Things (IoT). Both technologies are combined for goals such as detecting suspicious activity, cybercrime and fraud. Besides, a

Blockchain system that can handle massive data sets is compatible with ML approaches to data analysis and can increase data security [8]. Therefore, a huge variety of anomaly detection models are being designed and deployed by researchers for various Blockchains. However, one of the most difficult aspects of detecting fraud on the Blockchain is that it is anonymous [9].

Overall, it is necessary to note that anomaly detection is one of the important areas for protecting future blockchain networks and that a considerable amount of work is being undertaken on this subject from many views, which will be described in this paper. Ensemble approaches are prominent ways of increasing the prediction capacity of an ML model for anomaly detection. In theory, ensemble learning techniques use multiple classifier methods to improve experimental outcomes. Conventional methods that use a single classifier to perform predictive analysis are ineffective. Therefore, combining individual classifiers in an ensemble can produce higher accuracy values [112]. For instance, strategies include stacking, averaging, bagging, and boosting approaches [10].

This research focuses on the fundamentals of BT, ML classification, and the combined contribution of ML and Blockchain to detect irregularities utilizing ensemble techniques. To aid comprehension, the study is divided into three sections: (2) Blockchain principles, (3) an overview of ensemble learning classification, and (4) developing the ensemble learning method for anomaly detection in blockchain networks.

II. Blockchain Technology

A. Overview

Blockchain is presently one of the most promising technology trends, with great possibilities across many useful applications. It is basically a distributed and encrypted variation of a database, which can solve several difficulties connected to online security and trust. As a result, the Blockchain feature of securely and decentralized data management makes Blockchain known in the world of cryptocurrencies such as Bitcoin and Ether (Ethereum). Historically, the goal of producing interference-proof texts led to the development of a cryptographic hash formatting system for storing documents in a chain of blocks [11]. In this endeavour, hash-based cryptographic algorithms are used to store a collection of verified documents in Merkle tree format in each block [12]. Moreover, since it was invented and exploited in cryptocurrencies like Bitcoin, which was presented by Nakamoto [1] in 2008, this technology has become well-known. This has helped popularize Bitcoin as the first digital electronic payment mechanism that operates on a peer-to-peer (P2P) basis and in a decentralized ecosystem. The field of Blockchain has been divided into four categories: Private, Public, Hybrid, and Consortium. Fig. 1 depicts the categorization of Blockchain.

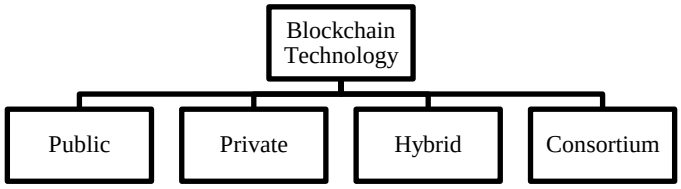


Fig. 1. Blockchain Classification.

A Public Blockchain is non-restrictive and permissionless [13]. This means anyone can do the mining process, and these transactions involve the addition of new blocks settled through a consensus mechanism. This concept has been fundamental to the existence of Bitcoin and other cryptocurrencies in the Distributed Ledger Technology (DLT) ecosystem [1]. As a result, the weakness of the centralized operating system faces challenges in terms of low-security level and no value of transparency (dependence on third parties). Regarding data storage, the DLT ecosystem stores data distributed across nodes linked by a blockchain network, as opposed to a centralized system that stores data in a single location. Technically, the consensus mechanism is an important algorithm in Blockchain operations to ensure that members joining a blockchain network agree on certain conditions before the ledger is updated. Proof of work (PoW) is a common consensus algorithm used in Public Blockchain environments. One of the benefits of this consensus is that as the number of miners grows, attacks can be reduced to 51 percent [16].

In contrast to the Public Blockchain, the Private Blockchain operates based on an organization through access granted only to be allowed to enter the network. Therefore, they are also called "permissioned blockchains" or "business blockchains" [17]. It has the same properties as a Public Blockchain that is distributed, decentralized, and operates in a P2P environment. Typically, a Private Blockchain is used in a network environment with a small organization compared to a Public Blockchain, where anyone has the right to enter a public network. The consensus algorithm used in the Private Blockchain (permissioned) is Practical Byzantine Fault Tolerance (PBFT).

Using both Public and Private Blockchain features in Blockchain development is necessary in the real world. As a result, a Blockchain ecosystem known as Hybrid Blockchain [18] has emerged. Elements from the Private Blockchain (permissioned) are employed in the enterprise context. On the other hand, a Public Blockchain is ideal for practice since the data requirements are open or public (permissionless). The addition of the participation of several organizations from a single organization so that the value of collaboration is higher in a Private Blockchain environment is termed a "blockchain consortium" [18]. It combines features of a Public and Private Blockchain and is very similar to a Hybrid Blockchain. An important goal is to eliminate access gaps limited to a single organization in a Private Blockchain environment.

B. Blockchain Architecture

In a decentralized ledger, all transactions in a Blockchain are stored in interconnected blocks. Each block contains a block header that stores critical information, including the timestamp, nonce, difficulty, block hash, and Merkle root tree, to keep these blocks related. This method guarantees the security of the data within the blocks, and the size of the witness determines the size of each block. One Bitcoin block, for example, is 1 MB in size [1]. Meanwhile, the Merkle tree employs the hash technique for each block transaction, as shown in Fig. 2. From an operational point of view, each block stores the address of the parent block or the previous block in the form of a hash value. This mechanism can help to identify the chain sequence between these blocks. Blocks generated in the early stages of blockchain network construction are termed "block genesis." To ensure the uniqueness of each block, the timestamp information is crucial to store the time differentiation generated on each block. For example, the current block has a more recent timestamp value than the timestamp of the previous block. This mechanism can prevent the occurrence of double-spending cases.

Blockchain environments, especially Bitcoin, are known for mining processes using pseudo-random numbers (nonce) and are used only once throughout the mining process. Note that it is difficult to keep the value of the difficulty level based on a threshold with a specific target. For example, the difficulty level rises when the number of transactions increases. As a result, block formation becomes increasingly complex (mining process) and slower. It also affects cyber attackers and greedy miners who want to take advantage of many transactions and slow the processing. The Merkle tree cryptographically manages the hash mechanism on transactions in blocks. This is described as a tree consisting of leaves as well as twigs. Conceptually, the hash in the brand tree is constructed based on a combination of left and right hashes to produce the parent hash. The generation of interconnected hashes forms a chain called a Blockchain. Therefore, an abnormality in the Merkle tree indicates something is happening in the chain, and appropriate action is taken immediately [19].

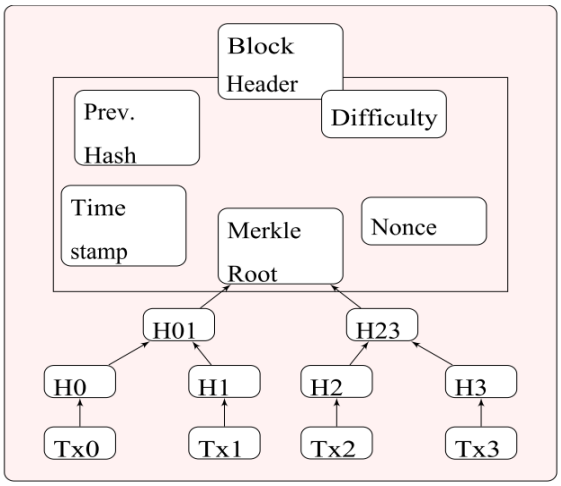


Fig. 2. Block Architecture [2].

C. Blockchain Layers

From the Blockchain Technology (BT) layer perspective, there are six layers in the blockchain network, as depicted in Fig. 3. The blockchain network contains several layers to execute specialized activities [20,21]. The data layer provides cryptographic techniques that store data in the hash, Merkle tree, and timestamp value forms in both on-chain (Blockchain) and off-chain (database) settings. The network layer manages all of the nodes in the blockchain network. At the network layer, this level of security and privacy is made sure to stay in place by a decentralized P2P environment. At the same time, transaction consistency is managed by consensus mechanisms located at the consensus layer. The mining process rewards successful miners. It is managed in the incentive layer. The condition of the smart contract in the Blockchain ecosystem is important to ensure that the security aspects are guaranteed, bug-free, and free from any vulnerabilities. Therefore, the smart contract programme is implemented at the contract layer. The application layer, which connects the end-user to the blockchain network, is the final layer. This layer comprises Blockchain applications (Decentralized Applications (DApps)) that were designed and constructed based on the business case in various sectors.

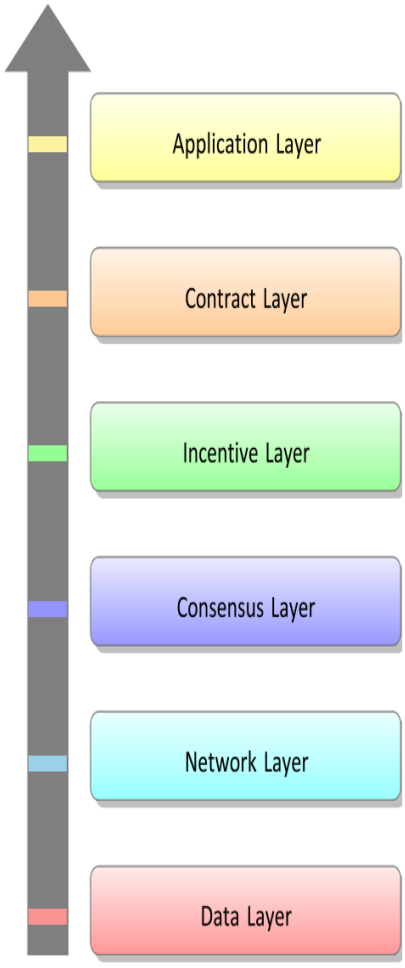


Fig. 3. Six Blockchain Layers [3].

D. Consensus Algorithm

The blockchain network must verify each software's ledger for consistency and clarity. This is performed by a few steps that follow certain rules during the transaction process. The verification process is carried out decentralized, with transactions completed in a distributed environment managed by P2P-connected nodes in the network. The approaches or algorithms utilized to reach a consensus are called consensus algorithms. Fig. 4 shows various widely used consensus algorithms, including Proof of Authority (PoA), PoW, Proof of Stake (PoS), and PBFT. Each node seeking to participate (mining) in the PoW consensus process must contribute resources by completing mathematical problem challenges [14]. This problem has a different level of difficulty. It is a consensus technique used in Bitcoin [1] and Ethereum [22]. In PoS, only one miner can generate new blocks from all participating nodes, while other miners waste incentives and energy resources on the blockchain network [15].

As a result, PoS works better when only those nodes can verify that their shareholders are permitted to participate. It avoids the circumstance where one node owns the network since no single node may hold 51 percent of the network's money[23]. As a result, PoS can efficiently cut energy consumption and reduce the number of miners, and the transaction speed can be boosted compared to PoW. It is critical to obtain mutual understanding in the PoA consensus to ensure the transaction is valid. The node's blocks must be certified by the verified node, and the process continues through the successive rounds as planned [24]. The PBFT consensus refers to a Byzantine military analogy that is difficult to reach consensus if no nodes have reached an agreement. The effort to reach this agreement based on the leaders with the most weight is called the PBFT consensus [25].

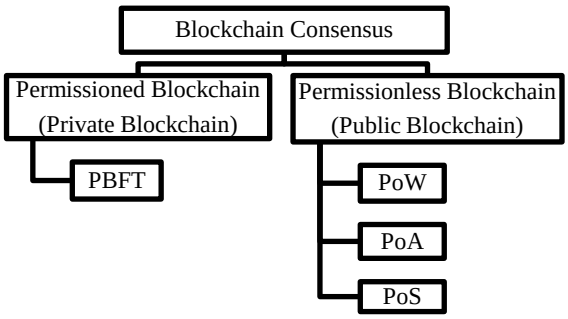


Fig. 4. Blockchain Consensus Algorithm.

E. Bitcoin

Cryptocurrency is one of the most extensively used Blockchain applications today and is used worldwide. Bitcoin and Ether (Ethereum) are two digital currencies commonly used in the crypto realm. Satoshi Nakamoto was the first to introduce Bitcoin, successfully solving the double-spending problem while introducing digital currency use [1]. The Blockchain controls each transaction using a cryptographic process based on hash values on input and output sets from an operational standpoint. Only one input transaction from the whole blockchain network is used to generate the output [26].

Aside from that, Blockchain is linked to a P2P ecosystem for transaction management and network ownership. The decentralization of Blockchain is a clear distinction between traditional databases and Blockchain. This implies that each network node is accountable for storing a copy of the ledger [19]. In the Bitcoin ecosystem, anyone can participate in the network. This feature is why Bitcoin is known by the term "incentive" or "reward" through the PoW consensus given to miners who successfully perform the mining process. As such, this Blockchain operates in a decentralized manner, which means it does not require a centralized body compared to traditional financial systems, which are centralized in nature. In this process, the miner gets paid a few Bitcoins after completing the operation.The mining process is secure because it involves hashed and encrypted transactions using the SHA-256 cryptographic technique. The popularity of Bitcoin as a Blockchain application for managing cryptocurrencies has prompted the development of several other crypto and DApps.

F. Ethereum

Buterin's paper [27] launched Ethereum and solved various problems with Bitcoin's scripting language. Ethereum had added transaction list and state information in the block header compared to before, which only contained information such as nonce, difficulty, and block number. A new state will be formed based on the previous state in the transaction list. The notable difference between Bitcoin and Ethereum is the cryptographic protocol used. Ethereum uses Keccak 256 bits while Bitcoin uses SHA-256. Thus, the header block in Ethereum consists of hashes for gas fee information, timestamp, parent block header, root state, and additional hashes for verification process purposes [28]. Ethereum provides a decentralized ecosystem for developers to develop products using the Solidity language and Ethereum Virtual Machine (EVM). The Solidity language is used to develop smart contract programmes based on business cases to be executed and converted to byte code in EVM [26].

G. Smart Contract

Historically, the idea of contract management has traditionally inspired the introduction of digital smart contracts by the founder of smart contracts, Szabo [29]. The main purpose of digital smart contracts is to automate traditional contract management. This smart contract is referred to as computer technology with the help of writing programme code to be implemented to automate the contract process. For operational purposes, smart contracts are integrated with Ethereum to be executed and stored in a decentralized ledger. Recently, the use of smart contracts has been widely used in conjunction with BT in various fields [61, 62]. Furthermore, the EVM environment and the Solidity programming language facilitate the development of smart contracts within Ethereum. This development has also attracted researchers to explore smart contracts on the Blockchain.

III. Ensemble Method

Machine Learning (ML) algorithms have been widely applied in both supervised learning and unsupervised learning situations to construct systems capable of making realistic decisions in light of past data. Numerous classification-based ensemble methods have been developed to boost the accuracy of supervised Learning Algorithms (LAs). Therefore, ensemble methods are prominent solutions for boosting the prediction capacity of an ML model. In the competition aspect, the ensemble approach has succeeded in several ML model competitions in which it has participated. For instance, the winner employed an ensemble method to create a robust collaborative filtering algorithm in the popular Netflix Competition [30]. Another example is Knowledge Discovery in Databases (KDD) 2009 when the winner also used ensemble methods [31].

Conceptually, the ensemble approach combines several trained individual classifiers to produce a new classifier. Typically, these individual classifiers are termed weak learners, and their ensemble combination aims to make this model stronger in terms of accuracy. However, among the challenges of using the original model individually is exposure to high variance and bias factors. Therefore, the ensemble strategy can reduce the bias and variance gaps to produce new combinations with better performance results, as illustrated in Fig. 5.

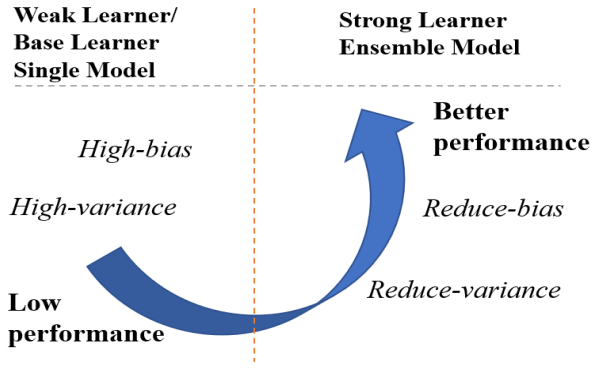


Fig. 5. Weak and Strong Learners.

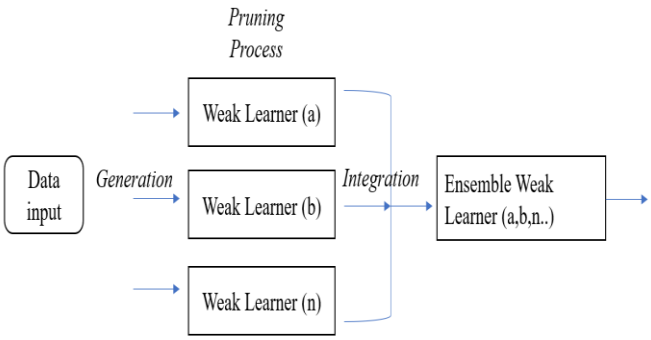


Fig. 6. Typical Process of the Ensemble Method.

With reference to Fig. 6, the process of ensemble generation from data input takes place in the first phase to produce weak learners. Next, the pruning and cleaning process is done for the weak learners. Finally, the combined integration of the weak learners is implemented in the last phase using the selected model. Past research has proven the ensemble approach successfully produces more accurate study results and lower false positive (FP) metrics than individual classifiers. The study also shows that popular ensemble strategies are stacking, bagging, and boosting. The authors [10] has described the ensemble as a variety of combined approaches consisting of the voting method, the averaging method, the stacking method, the bagging method, and the boosting method. According to [32], the ensemble approach can address the shortcomings of traditional ML, such as mathematical, computational, and representation problems. Fig. 7 depicts the ensemble learning methodology and methods. Moreover, the authors explain an ensemble as a model that incorporates the results from numerous other models to remedy the flaws of every situation. Most of this strategy's options can be classified as bagging or boosting [33]. In the averaging approaches, the authors [34] tests with different alternatives of anomaly detection models. The authors believe that choosing a simple average score between different algorithms is a simple and successful solution. Apart from that, the authors define combining the multiple models as needed because they address the problem from diverse aspects [34]. Using ensemble learning, the combination of Random Forest (RF), Extra Trees, and Bagging classifier demonstrated a possible performance by gaining the predictions based on averaging the probabilities derived from these methods [35]. The authors [36] describe how the results generated from the individual classifiers have enhanced their capabilities and have shown improved performance on the study results through the ensemble method. Meanwhile, the study by [113] used a Deep Learning (DL) approach to produce prediction analysis with an ensemble combination for a single classifier based on medical datasets. The study results show that the ensemble technique produces high accuracy values compared to the individual classifiers. Nowadays, more studies lead to new methods or techniques for model optimization compared to before, which is more to developing new models. Among them is a study conducted by the authors [114] using ensemble techniques to develop a new model optimization method for the prediction of taxological applications. The experimental results in this study show that the ensemble technique produces better results than the single classifier.

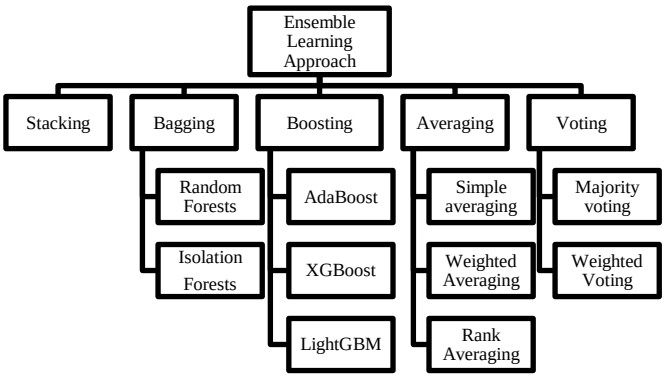


Fig. 7. Ensemble Learning Approach.

A. Voting

Voting is the easiest ensemble procedure. Among the main techniques of the voting ensemble is the majority voting ensemble, sometimes called the max voting ensemble. This is an ensemble strategy that combines multiple different types of individual classifiers. The desire to increase performance from individual models is an essential strategy. For classification and regression, ensemble voting might be used. The mean value of the forecast is derived using the regression approach. In the classification approach, labelling is based on the number of prediction outcomes tagged and the majority of votes. In practice, ensemble voting is appropriate when all individual models show good performance. Fig. 8 illustrates a voting ensemble learning illustration. In a study by [37], the majority voting-based ensemble model method was used. The results successfully detected network traffic as if there had been an attack on the Intrusion Detection System (IDS). In this research, the authors [37] mentioned that many classifiers were employed for training and testing, and final findings were attained utilizing the voting approach. Aside from the majority vote approach, the researchers chose to perform the investigation using the weighted voting method. Repeated calculations on the model prediction are used in the weighted voting method to produce a favourable result from the standpoint of the ballot weights. In the current work, weighted majority voting was used to categorize the data, where Particle Swarm Optimization (PSO) was employed for allocating weights to several classifiers [37].

B. Averaging

Using the averaging method, the simplest strategy for making predictions from dataset inputs is based on average values. In general, this method generates a better regression model and reduces overfitting. Nevertheless, this averaging variant is slightly modified to be a weighted average model. The prediction generated from this model is calculated based on the average value generated from the multiplication operation by the weights on each model. Rank averaging is the process of allocating ranks to individual models based on the weight to be assigned to each model. The method of averaging and determining the maximum score is one of the combination methods that can be used. The findings of the pilot experiment reveal that weighted averaging has been utilized to normalize the anomaly scores. This is done before combining the method to balance the results of unbalanced for different algorithms with different datasets [38]. The weighted average is the result of the study's final output based on the method of grouping the list of scores and assigning a weighting value that is inversely proportionate to the group size possessed by each list of scores, according to [39]. Fig. 9 illustrates the average ensemble learning demonstration.

C. Stacking

Stacking, or layered generalization, is an alternative way of integrating numerous models. In the stacking technique, various individual (multiple) models have been integrated. Among them are logistic regression (LR), Naïve Bayes (NB), and Decision Tree (DT). The learning approach of stacking is for merging the expectations of several classification models into a single meta-classifier [31]. Meanwhile, the authors [40] explained that stacking techniques in the ML approach could produce a more powerful model. This is implemented through training on datasets on individual models to improve accuracy. Basically, the stacking method uses the predictions made by a single model to make another model.

From an operational point of view, the stacking technique is carried out sequentially. The process begins by training several selected individual models using a dataset sample. Subsequently, the production probability results from each individual model go through a fine-tuned process before being combined into a final model. This procedure is performed repeatedly depending on the number of stacking layers you want to use. Finally, the final output is formed based on the final output generated by several individual models in the last layer. Therefore, the individual models generated at this end layer are known as meta-classifiers. According to [41], the learning output at the base layer determines the final output produced by the stacking method. Fig. 10 depicts the usual two-layer stacking modelling approach.

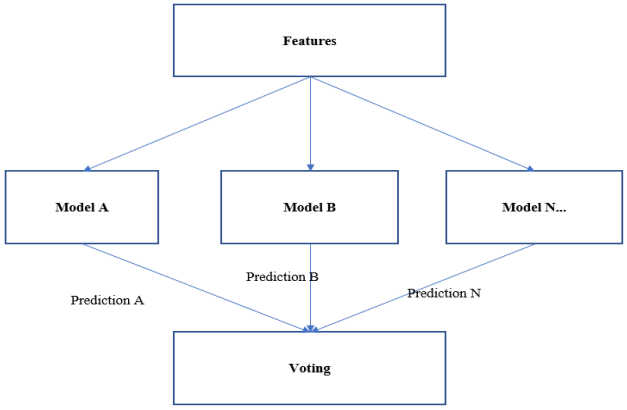


Fig. 8. Voting Ensemble Learning Illustration.

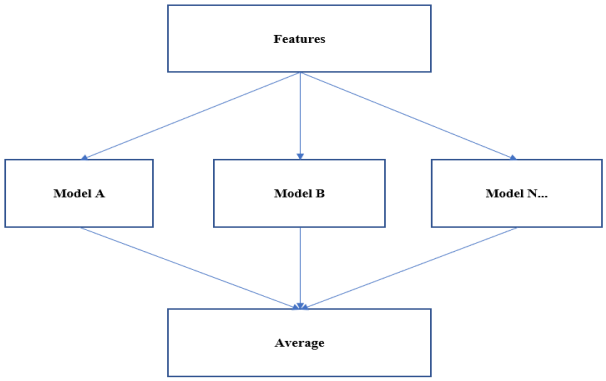


Fig. 9. Averaging Ensemble Learning Illustration.

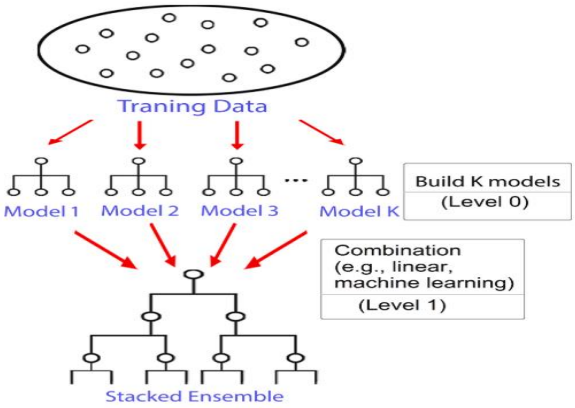


Fig. 10. Stacking Ensemble Learning Illustration[6].

D. Bagging

The bootstrap aggregating (bagging) was first described in [43]. It is one of the simplest ensemble approaches and is best suited for issues involving small training datasets. Sequential and parallel ensemble methods are the two predominant paradigms for constructing ensemble models. Technically, various series of datasets are formed through random extraction from samples of the original data set, and these data sets are used to train different models. Then, voting is used to aggregate the results of the models to form a single output. Bagging is used in regression and classification to improve the precision of ML algorithms. Besides, bagging also utilizes the most prevalent techniques for combining the outputs of base learners, namely averaging for regression issues and voting for classification tasks. Among the algorithms commonly used in the bagging technique is the DT. According to [44], this algorithm can be compatible with weak models and have high variance. However, apart from the DT, other model classifications such as K-Nearest Neighbour (KNN) and NB are also used in the bagging technique. Furthermore, creating a model using a simple method that incorporates large and complex data is impossible. Consequently, bagging approaches are ideal for managing both high-dimensional and large-capacity data. Fig. 11 depicts an illustration of the Bagging algorithm procedure.

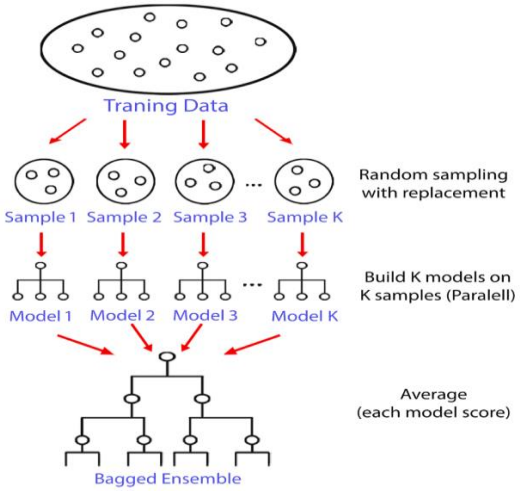


Fig. 11. Bagging Ensemble Learning Illustration[6].

1) Random Forest (RF): RF was introduced in 2002 by Breiman. The Random Forest is, as its name suggests, a forest comprised of numerous trees. In general, RF (Tree-Based) use a DT as an individual model, which generates a set of random parameters as the value of dependence on each tree. Similar to other ensemble algorithms, RF produces predictions by combining numerous separate models. Basically, the RF procedure consists of multiple steps. First, bootstrap samples were randomly generated from the dataset. Then, the prediction results of each tree will be obtained from the construction of the DT based on the data sample. Lastly is the implementation in the voting phase to produce the final output. In this last phase, the model that gives the most accurate prediction results will be selected [45].

2) Isolation Forest (IF): The Isolation Forest (IF) algorithm was first proposed in 2008 [46]. Like any other tree ensemble method, this approach is based on DT. It operates on the premise that an individual who is easier to distinguish from others in a random sub dataset of the feature space must be an outlier. It begins by drawing a random sample from the dataset and selecting a random dimension. Correspondingly, a random value within the range of that dimension is selected to precisely divide the sample into two pieces. Next, the root node of a tree is built using the selected dimension and splitting point. Further nodes are produced recursively for subsamples until a subdivision is impossible or an arbitrary tree depth is attained. In this tree, a point closer to the root node correlates to a situation more likely to be isolated. Nevertheless, this could be due to random chance. Therefore, the entire tree generation technique is repeated for additional samples until the necessary number of trees is achieved. Note that the anomaly score is computed using the mean traversal path length of the trees. The authors of [46] claim that their algorithm is superior to other alternatives for addressing masking difficulties (clusters of anomalies) and swamping problems (mistakenly identifying normal situations as being surrounded by anomalies).

E. Boosting

Boosting is a strategy for enhancing the performance and accuracy of the ML approach by transforming weak base learners into strong ones [47] as shown in Fig. 12. The fundamental premise of the boosting strategy is to sequentially add new models to the ensemble. In general, the boosting technique generates a sample of training data randomly with the replacement of the main dataset sequentially. In this procedure, a sequence of models is learned. The process begins by providing training on the weak model using a training dataset to produce a second model after fixing the weaknesses in the first model. Subsequently, a third model was produced that overcame the weaknesses of the previous two models. This process will continue until all the mistakes are fixed and the final model is made. Last, a technique weighted majority voting was used to build the final model from the weak model [48,49]. Boosting techniques have been proven to increase accuracy and reduce bias and variance. Among the algorithms widely used in boosting techniques are

Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosted Machine (LightGBM).

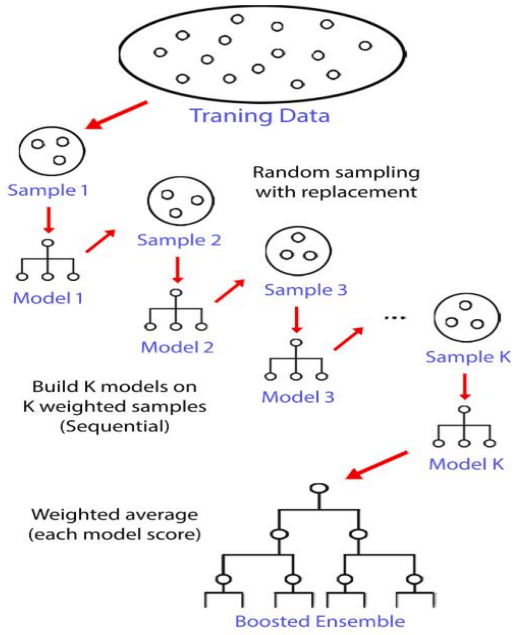


Fig. 12. Boosting Ensemble Learning Illustration [6].

1) Adaptive Boosting (AdaBoost): AdaBoost was the first truly successful binary classification boosting method. It was originally referred to by its inventors as AdaBoost.M1 [50]. Recently, it has been referred to as "discrete AdaBoost" because it is utilized for classification instead of regression. AdaBoost, like other approaches, may be used to increase the performance of any ML model and can be used for learners with low intelligence. This strategy works by turning weak learners into strong ones by getting rid of them by correcting their mistakes over and over again iteratively. The weighted training dataset is used to train weak learners in succession. Subsequently, numerous weak learners are joined to become a single powerful learner. Finally, the weight voting method on the weaker model was used to determine the stronger final model [50]. Besides, one-level DT is the best-suited and, thus, the most popular algorithm employed with AdaBoost. Since these trees are so short and contain only one classification decision, they are often referred to as decision stumps.

2) Extreme Gradient Boosting (XGBoost): Extreme Gradient Boosting, or XGBoost, is a scalable ML approach for tree boosting that was presented by Chen and Guestrin [51]. XGBoost is a gradient boosting-based model that uses additional boosting strategies to produce predictions more accurately compared to other gradient boosting models [52]. Therefore, the advantages of this technique have been acknowledged in various fields of ML and data science. For example, a total of 17 winners used the XGBoost technique out of a total of 29 winners to complete one solution contest as well as be featured in the Kaggle blog [53]. XGBoost uses the advantages of boosted tree algorithms to produce accurate and scalable boosting gradients. Moreover, XGBoost has been designed with fast computer processing and improved ML model performance in mind. In general, XGBoost works in parallel to generate trees. This process is implemented level by level to produce predictions on each iteration from weak learners. As a result, each of these iterations can improve the errors of their predecessors. The final result of prediction with a combination of individual models and these mechanisms is the same as with other ensemble approaches.

3) Light Gradient Boosted Machine (LightGBM): LightGBM, or Light Gradient Boosted Machine, was described by Guolin Ke et al. in 2017 [54]. LightGBM is a gradient boosting implementation aimed to be efficient and possibly more successful than previous gradient boosting implementations. According to the authors [54], the solution includes two main concepts: 1) Gradient-based One-Side Sampling (GOSS); and 2) Exclusive Feature Bundling (EFB). GOSS is a variation on the gradient boosting approach that prioritizes training samples that provide a greater gradient, accelerating learning and minimizing the method's computing complexity. In contrast, EFB is a method for combining sparse (mainly zero) mutually exclusive features, such as one-hot encoded categorical variable inputs. Consequently, this is a form of automatic feature selection. Through this concept, LightGBM has adapted a tree algorithm capable of producing high performance, classification, ranking, and various tasks in ML. Besides, LightGBM is a fast, more efficient, less memory-intensive, more accurate than any other boosting algorithm, compatible with large datasets, and gradient boosting framework. Normally, the DT through the boosting method is determined based on their level or depth. Nevertheless, this approach differs from LightGBM, which divides the tree based on the optimal leaf. Therefore, this approach provides a high level of accuracy by minimizing the level of loss and is an achievement that is rarely achieved by any existing booster algorithm.

IV. Ensemble Anomaly Detection in Blockchain

Nowadays, the development of Blockchain Technology (BT) is not just focused on the world of cryptocurrency but its expansion to Decentralized Applications (DApps) in various fields. Following this, the features available in BT have provided advantages in terms of transparency, immutability, enhanced security level, fast transactions, and high privacy. As a result, we see many applications that use BT in various sectors, namely finance, supply chain, halal products, pharmaceuticals, education, government, etc. In cryptocurrency, Bitcoin and Ethereum are the most popular and widely used applications due to their high market capitalization and trading volume. Apart from that, Bitcoin constitutes about 39.53 percent of the market's entire value [55]. At the same time, Ether is the second-biggest cryptocurrency [3]. Meanwhile, Ethereum is the largest and most widely used decentralized Blockchain platform for smart contract adaptation. The widespread use of Bitcoin, as well as Ethereum, has given rise to some critical issues in the aspects of cybercrime and security. As a result, many have become

victims of various frauds, such as phishing and Ponzi Schemes, after detecting more than 10 percent of Initial Coin Offering (ICO) on Ethereum. Generally, the Ethereum blockchain network is a public distributed ledger with around 1.158 million daily transactions [56] and is categorized as big data. Therefore, manually combing through all of these transactions to find any transactions suspected of exhibiting unusual characteristics would be impracticable and interminable. Based on this scenario, Machine Learning (ML) algorithms would help differentiate between transactions that exhibit normal and abnormal behavior among user accounts by learning the attributes that correspond to either normal or abnormal conduct.Therefore, an approach to detecting transactions that show abnormalities was introduced, known as the abnormal detection method. Nowadays, this method is increasingly used in various fields to detect patterns of abnormalities, especially its role in the Blockchain ecosystem. The detection model developed using the ML model helps detect and predict the initial attacks on the blockchain network. Fig. 13 offers data visualization for normal and anomalous transactions to better understand anomaly transactions. Oddities or unusual occurrences have the same meaning as deviations, noise, novelties, exceptions, and outliers [7]. Clearly, the combination of Blockchain and ML technology positively benefits both parties, as shown in Fig. 14. The Blockchain ecosystem is known for its overly large data storage nature and can be declared big data. There is also data from external sources such as smart devices, the Internet of Things (IoT), and external applications that store data in a database (off-chain). Thus, data from various sources is analyzed using ML techniques to produce analytical dashboards, predictions, visualizations, and others that can help with planning, monitoring, and decisions.

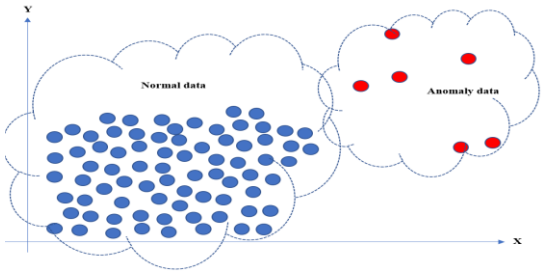


Fig. 13. Data Visualization for Normal & Anomaly.

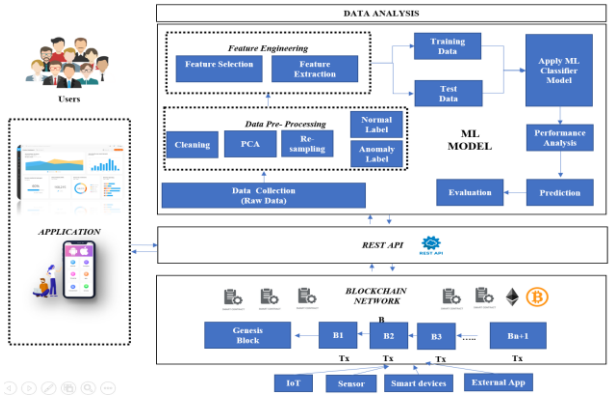


Fig. 14. Connection between Blockchain and ML.

In earlier study, numerous ML algorithms have been applied in supervised [57] and unsupervised learning [58] for anomaly detection in blockchain networks. Random Forest (RF) [59], Decision Tree (DT) (j48) [60], Extreme Gradient Boosting (XGBoost) [61], Adaptive Boosting (AdaBoost) [62], secureSVM [63], Light Gradient Boosted Machine (LightGBM) [64], K-Nearest Neighbour (KNN) [65], Support Vector Machines (SVM) [66], Naïve Bayes (NB) [67] and Isolation Forest (IF) [68] are examples of supervised learning models. Among the models in unsupervised learning that have been utilized are One Class Support Vector Machine (OCSVM) [69], K-means [70], Density Based Spatial Clustering of Application with Noise (DBSCAN) [71] and Long Short Term Memory (LSTM) [72]. This article evaluates the ensemble learning method for detecting anomalous or criminal transactions in blockchain networks. Ensemble learning gave good results and great performance in the experiments for recognizing malicious Ethereum entities [73]. Moreover, the authors execute ensemble learning, a mixture of ML predictors that wins over other classical learning approaches at predicting licit and illegitimate transactions. In the experiment, ensemble learning can be characterized as a classification method based on an average probability ensemble constructed from the collection of best-performing supervised learning methods employed in our experiment [35]. However, individual classifiers are troublesome for processing high-complexity data, according to [74] research. Consequently, this issue has been handled by developing a classification model utilizing the ensemble approach. In a Proof of Concept (PoC) development project for the decentralized unmanned aerial vehicle (UAV), the ensemble stacking method was applied to a variety of individual models to assess its predictive accuracy [75].The completed literature evaluation led to the classification of prior research articles about the addressed applications published from 2017–2022. Publications were divided into four aspects: anomaly detection in cybercrime (see Table I), security (see Table II), information processing (see Table III) and smart devices (see Table IV).

A. An Anomaly in the Aspect of Cybercrime

Cybercrime means using computers, tools or materials with the intent to do illegal things [76]. BT's openness, transparency, and immutability have prompted malicious parties to commit criminal activities. Most cyberattacks are performed for financial benefits. In the cryptocurrency era, hackers are prompted to get their ransoms in cryptocurrencies, as it provides the advantage of anonymity and easy transfer across countries. Therefore, among the effective methods is to use ML techniques to detect abnormalities in blockchain network transactions. Many previous studies have reported detecting transaction abnormalities using the approach of the abnormality detection method. Thus, in this review, we identified 31 publications that apply the cybercrime aspect in the selected papers, as shown in Table I. Referring to Table I, cybercrime aspects are categorized according to the type of application case, namely smart contracts, illicit transactions, scams (pump and dump), fraud detection, ransomware, Ponzi Schemes, money laundering, High Yield Investment Program (HYIP), and phishing, as shown in Fig. 15.

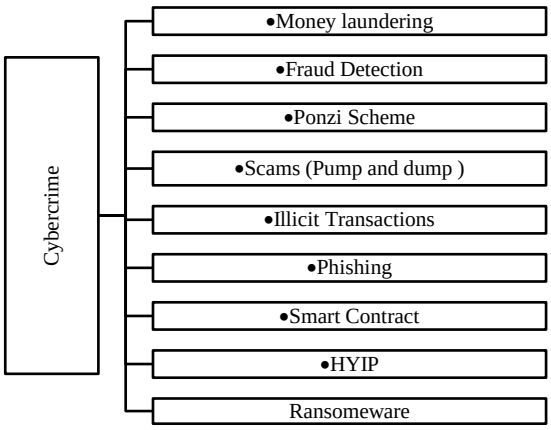


Fig. 15. Classification of Application in the Cybercrime Aspect.

As indicated in Table I, an RF was the most commonly utilized ensemble model (based learner) in the chosen research publications. In addition, 20 research publications utilized bagging as an ensemble approach, and 11 research papers embraced the boosting method. Furthermore, we uncovered 24 research articles utilized in Bitcoin and Ethereum. Since developing an ML model relies on the dataset, we analyzed the data source of ML models for anomaly detection applied in the selected research publications. The analysis of data sources has shown that 30 different types of data sets were used in the experiment. In earlier investigations, it was observed that there are numerous ways employed in the ensemble learning method. Among them is combining ensemble approaches or tactics to produce a good result. The review papers describe numerous techniques for this hybrid scenario, including bagging with voting, bagging with averaging, bagging with boosting and bagging with stacking.

The authors [77] suggested a pre-encryption detection algorithm (PEDA) that seeks to identify ransomware using an ML approach to assess and categorize ransomware using the bagging and voting (majority voting) ensemble learning technique. This research was conducted in Phase 1 and Phase 2 using a dataset created by Resilient Information System Security (RISS) from Imperial College, London. Nevertheless, the focus of this study is the focus on Learning Algorithm (LA) implemented in Phase 1. In general, LA works through an ensemble DT approach. First, the simulations of the LA model were implemented using the Application Programme Interface (API) data generated by suspicious software for inspection. Then, performance measurement analysis was performed by comparing the LA model with three other models, namely NB, RF, and ensemble techniques (RF and NB). Finally, this model was selected using the majority voting method. The results of this experiment have shown that the LA model produces better performance compared to the individual models' RF, NB and the ensemble models (RF and NB). Measurement metrics use detection rate (DR), False Positive Rate (FPR), Under Area the ROC Curve (AUC), and test error values.

TABLE I. SUMMARY OF PREVIOUS RESEARCH USING ENSEMBLE METHOD IN CYBERCRIME ASPECT

Ref.	Year	Blockchain Application	Application	Ensemble method applied	Model/Based learners	Tools/Dataset
[79]	2017	Bitcoin	Fraud detection	Bagging	Random Forest	Public Dataset
[80]	2017	Ripple	Anomaly detection	Averaging	One Class SVM, Gaussian Mixture Models, Isolation Forest	Ripple Transaction dataset
[81]	2017	Bitcoin	HYIP	Bagging, Boosting	Random Forest, XGBoost	Public Dataset
[82]	2018	Bitcoin	Ponzi Scheme	Bagging	Random Forest	Public Dataset Reddit, Bitcointalk.org
[83]	2018	Cryptocurrency	pump and dump scams	Bagging	Random forest	Telegram API Twitter API Crypto Market Data
[84]	2018	Ethereum	Ponzi Scheme	Boosting	XGBoost	Etherscan API/Real Data
[82]	2018	Bitcoin	Ponzi Scheme	Bagging	Random Forest	blockchain.info public dataset (bitcoinponzi)
[85]	2018	Bitcoin	De-Anonymising Entity	Boosting	Gradient Boosting	Chainalysis
[53]	2019	Ethereum	Fraudulent Accounts	Bagging	Random Forest	Etherscan API/Real Data
[86]	2019	Cryptocurrency	Anomalous transactions	Bagging	Random Forest	Etherscan API/Real Data Binance
[86]	2019	Cryptocurrency	pump and dump scams	Boosting	XGBoost	Binance Telegram Data
[87]	2019	Ethereum	Ponzi Scheme	Bagging	Random Forest	Etherscan API/Real Data
[88]	2019	Bitcoin	HYIP	Bagging	Random Forest	WalletExplorer Blockchain.info Xapo.com
[77]	2019	Bitcoin	Crypto-ransomware	Bagging, Voting	Naive Bayes, Random Forest	RISS dataset API Cuckoo Sandbox

						SQL database
[60]	2020	Bitcoin	Illicit entities	Bagging	Tree-based	VJTI Blockchain lab
[89]	2020	Ethereum	Illegal activity	Boosting	XGBoost	Etherscamdb Etherscan API
[90]	2020	Bitcoin	Money Laundering	Bagging	Random Forest	Elliptic
[91]	2020	Ethereum	Fraudulent Behaviour	Bagging	Random Forest	etherscamdb.info
[35]	2020	Bitcoin	Anti-Money Laundering (AML)	Bagging Averaging	Random Forest, Extra Trees, and Bagging classifier	Elliptic
[92]	2020	Ethereum	Honeypot Smart Contract	Boosting	LightGBM	Honeybadger,Ethereum Client,Parity Client
[93]	2020	Ethereum	Ponzi Scheme	Boosting	Ordered Boosting	bitcointalk.org,Google BigQuery,PonziTect
[94]	2020	Bitcoin	Fraudulent Transactions	Bagging	Random Forest	Kaggle
[86]	2020	Cryptocurrency	Fraudulent Transactions	Bagging	Random Forest	Etherscan API
[95]	2020	Cryptocurrency	pump and dump scams	Bagging	Random Forest	Telegram, Twitter, Reddit, BitcoinTalk
[59]	2021	Ethereum	Fraudulent detection	Bagging	Random Forest	Kaggle
[96]	2021	Bitcoin	Fraud Transactions	Bagging	Random Forest	Bitcointalk,bitcoin public dataset
[97]	2021	Ethereum	Fraudulent Detection	Bagging	Random Forest	Google BigQuery Github
[62]	2021	Ethereum	Phishing	Boosting	AdaBoost	Etherscan API
[78]	2021	Ethereum	Fraudulent Transactions	Bagging, Boosting	Random Forest, Adaboost, SVM	node2vec
[98]	2021	Ethereum	Vulnerability Detection	Boosting	XGBoost	Etherscan API
[74]	2022	Cryptocurrency	Anomaly Detection	Boosting, Stacking	SVM, KNN Logistic, DT, MLP	Kaggle

Adapting ensemble techniques has also worked well in networking, where they have been used to predict both licit and illicit transactions [35]. In this experiment, the approach of bagging with averaging technique has been applied to anticipate licit and criminal transactions in the blockchain network. The proposed approach of an ensemble (RF, Extra Trees, and Bagging classifiers) has fared the best with a comparison of RF, Multilayer Perceptron (MLP), and Logistic Regression (LR). In an average probability ensemble, the classification is done by employing numerous pre-trained ML models. The final predictions are formed by averaging the summation of the prediction probabilities received from the LAs. Note that the results demonstrate that ensemble learning is able to execute classification with an accuracy (98.13 percent) and F1 score (83.36 percent) to forecast licit and illegal transactions.

The authors [78] gives a comprehensive evaluation of different supervised ML algorithms, such as bagging models (RF), boosting models (AdaBoost), and others, to prevent fraud. This research concluded that utilizing AdaBoost and RF classifier produced the best performance result among the other seven algorithms.

Feature selection in the ensemble approach plays an important role in producing better results. This has been demonstrated by [74], who conducted studies on the use of feature selection and without feature selection. This simulation is performed by comparing the use of feature selection with that without feature selection in the ensemble classifier (boosting, stacking). The final results have shown that there is an increase in the value of F-Score (7 to 9 percent) and accuracy (2 to 3 percent).

B. An Anomaly in the Aspect of Security

BT does not guarantee freedom from security issues. Therefore, there is a need to establish risk management through a comprehensive cyber security framework and undergo security assessment services to protect against attacks and abuse by hackers. This security issue has been researched and has found a total of 31 research papers involved in the study on the aspect of security, as shown in Table II. This in-depth study uses ensemble techniques to find anomalous transactions in a blockchain network. According to Table II, security elements are largely split into backdoor assaults, vulnerability identification, crypto-jacking, under-priced Denial of Service (DoS) attacks, intrusion detection, miner detection, malware, cybersecurity framework, protection of private information, botnet and malicious account detection, and so on, as shown in Fig. 16.

As shown in Table II, an RF was the most commonly utilized ensemble model (based learner) in the selected research publications. In addition, four research publications utilized bagging as an ensemble approach, two research papers adopted the stacking method, and 1 research study applied to boost and to vote. Moreover, we identified four research publications that have been used in Ethereum. The utilization of datasets is the crucial component of ML model construction. Consequently, this study's analysis considers the datasets utilized in prior studies. As a consequence, it was determined that the selected research utilized five distinct types of data sets. In the ensemble approach, a combination of several ensemble (hybrid) techniques is used to achieve better performance in the study. Among them are: In reviewing investigations for security considerations, it was determined that two research publications used combined ensemble methods or strategies to achieve a decent outcome. In addition, there is one research paper that utilized the stacking with boosting strategy and one paper that used the bagging with the voting approach. The authors [73] offered strategies for detecting malicious entities that employ versions of RF, SVM, LR, and ensemble methods with stacking and boosting (AdaBoost Classifier). With an average F1 score of 0.996, the study's findings demonstrate that the ensemble technique yields effective outcomes. This study's strategy is to establish a framework for identifying entities that potentially do harm to blockchain networks.

The conventional Exploratory Data Analysis (EDA) methodology is implemented via data collection, feature extraction, model training, model testing, and final outcomes evaluation to achieve this objective. The study's results also demonstrated that feature extraction is an effective strategy for achieving positive outcomes.The research on under-priced DoS assaults was proposed by the authors [99]. In this study, the simulation method is implemented on the transaction using several input features, namely pending time, value, gas price, and gas. Several ML models were used in this study, such as NB, SVM, KNN, RF, and DT. While the voting technique, which consists of two criteria, namely majority vote (hard) and average confidence (soft), is practiced. This study concluded that the experimental results had shown good performance in detecting under-priced DoS attacks. Conventional UAVs generally depend upon the centralized server to execute data processing with complicated ML techniques. In reality, all classic cyberattacks are relevant to data transmission and storage in UAVs. In this regard, [75] proposes to boost the performance of UAVs with a decentralized ML architecture based on Blockchain. In general, UAV or drone technology uses centralized data processing technology. Unlike a decentralized Blockchain, it is vulnerable to cyberattacks on storage and transactions. Thus, [75] has studied this matter by providing added value using the ML method in Blockchain applications to generate prediction analysis and improve UAV performance. This study also aims to prove that the centralized ML model approach has improved resource utilization and overhead performance. Following this, the decentralization of the ML model is a wise move to produce high-quality forecasting. Therefore, this study conducted two experiments using stacking techniques and without stacking. This study found that using PoC stacking has made forecasting analysis more accurate.

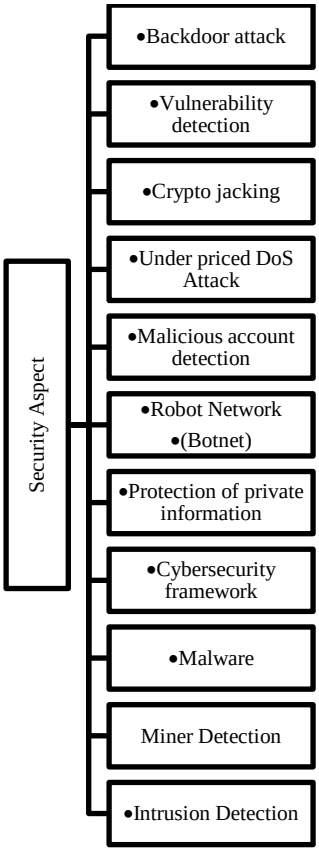


Fig. 16. Classification of the Security Aspect.

C. An Anomaly in the Aspect of Information Processing

Information processing is capturing, recording, organizing, retrieving, displaying, and disseminating information. The word has often been applied to computer-based activities in recent years. In this part, we identified 31 papers that apply the information processing characteristics in the selected publications. The list of these applications shows in Table III. According to Table III and Fig. 17, information processing components are primarily categorized as Blockchain simulator, performance testing, network traffic, social media, data analysis, address identification, performance testing, transaction clustering and behavioural pattern

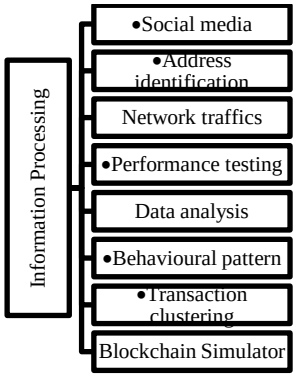


Fig. 17. Classification of Information Processing Aspect.

TABLE II. SUMMARY OF PREVIOUS RESEARCH USING ENSEMBLE METHOD IN THE SECURITY ASPECT

Ref.	Year	Blockchain Application	Application	Ensemble method applied	Model/Based learners	Tools/Dataset
[67]	2019	Ethereum	Vulnerability detection	Bagging	Random Forest	Etherscan API
[73]	2020	Ethereum	Malicious Transaction	Stacking, Boosting	Random Forest, Stacking Classifier, AdaBoost	Ethereum Client, Etherscan API
[100]	2020	Blockchain-based	Crypto-jacking	Bagging	Random Forest	VirusTotal
[101]	2021	Ethereum	Malicious Account	Bagging	Tree-based	Etherscan API
[99]	2021	Ethereum	Under-priced DoS attack	Bagging,Voting	DT, Random Forest,KNN, SVM	Ganache
[75]	2021	Blockchain-based	intrusion detection	Stacking	KNN, NB, SGD, Onevsrest, Logreg	KDD99 attack dataset

TABLE III. SUMMARY OF PREVIOUS RESEARCH USING ENSEMBLE METHOD IN INFORMATION PROCESSING ASPECT

Ref.	Year	Blockchain Application	Application	Ensemble method applied	Model / Based learners	Tools/Dataset
[64]	2019	Bitcoin	Address Identification	Boosting	LightGBM	WalletExplorer, Blockchain.info, BitcoinTalk
[103]	2019	Bitcoin	Network Traffic	Bagging	Random Forest	WalletExplorer
[102]	2019	Bitcoin	Data Analysis	Stacking	Random Forest Gradient Boosting (GB)	WalletExplorer

As indicated in Table III, there are three research articles, and the most commonly employed ensemble model (based learner) in the selected research papers was an RF. In addition, one research paper utilized bagging as an ensemble approach, one research paper adopted the stacking method, and one research paper applied to boosting method. Furthermore, we uncovered three scientific publications that have been utilized in Bitcoin. Finally, note that the development of the ML model depends on dataset input. Thus, this analysis has looked at three different types of data sources used in selected studies. In this study, the authors in [102] employs cascading ML principles—a sort of ensemble learning employing stacking techniques. This study's simulations utilized weak classifiers, GB and RF. As a result, the ensemble stacking method yielded effective classification outcomes based on F1-score, recall, and accuracy values.

The voting-based method developed by the authors [103] aims to improve the level of tracking of Bitcoin performance by labeling addresses controlled by the same user. This study uses Bitcoin datasets taken from previous study publications [104,81] and WalletExplorer. Through simulations on Bitcoin addresses of 200K, we found that the voting method produces better results than the non-voting method in terms of F1 score, recall, and precision. Labeling using supervised learning methods was used to develop a model classification for detecting anomalies in Bitcoin addresses [64]. Therefore, this experiment was conducted using eight main classifiers, namely LightGBM, XGBoost, NN, AdaBoost, RF, SVM, Perceptron, and LR. The experiment showed that the LightGBM classifier produced the best results with a micro/macro score value of F1 (97 percent/86 percent).

D. An Anomaly in the Aspect of Smart Devices

Smart devices are generally IoT gadgets with support for Internet connectivity. They can interact with other devices over the Internet and offer remote access to a user for operating the device as per their needs. In this section, we selected three papers exploring smart device applications. The list of these applications is shown in Table IV. According to Table IV, smart device characteristics are largely grouped, as illustrated in Fig. 18.

As indicated in Table IV, there are two research articles, and the most often employed ensemble model (based learner) in the selected research papers was XGBoost and Adaboost. In addition, two research publications utilized boosting. Furthermore, we located 1 research paper used in the Blockchain-based Blockchain simulator. From the perspective of datasets, the study has identified four distinct dataset categories used in the selected studies. This is because the ML model to be constructed is dependent on the dataset used.

The authors in [61] describe the design and architecture of our Blockchain simulator, BlockEval, which simulates the behaviour of concurrent activities in a real-life Blockchain system. This research confirmed the correctness of our simulator by comparing it with an independent model constructed using genuine Bitcoin transaction data. XGBoost is a non-parametric supervised LA used for classification and regression. The goal value is anticipated by learning simple decision rules inferred from data attributes. Simulation results have been drawn up to 2000 nodes, which have been checked against actual Bitcoin data. However, there is a scope of enhancement to both the simulator and the validation architecture. For instance, adding propagation latency data with a suitable variance will increase the accuracy of simulation findings. IoT-related research has been undertaken

by [42], concentrating on data integrity and security. An important thing to perform is to discover irregularities in data transactions using ML approaches. Hence, the IoTID20 dataset, consisting of 80 characteristics (62578 records), was utilized for training the model to be constructed. This study was conducted by taking 15 traits designated as normal and abnormal. During this investigation, different model classifications were trained based on measurement parameters such as F1 score, recall, precision, and accuracy. The experimental results reveal that the AdaBoost and RF algorithms provide similar results and are among the highest classifiers with good performance.

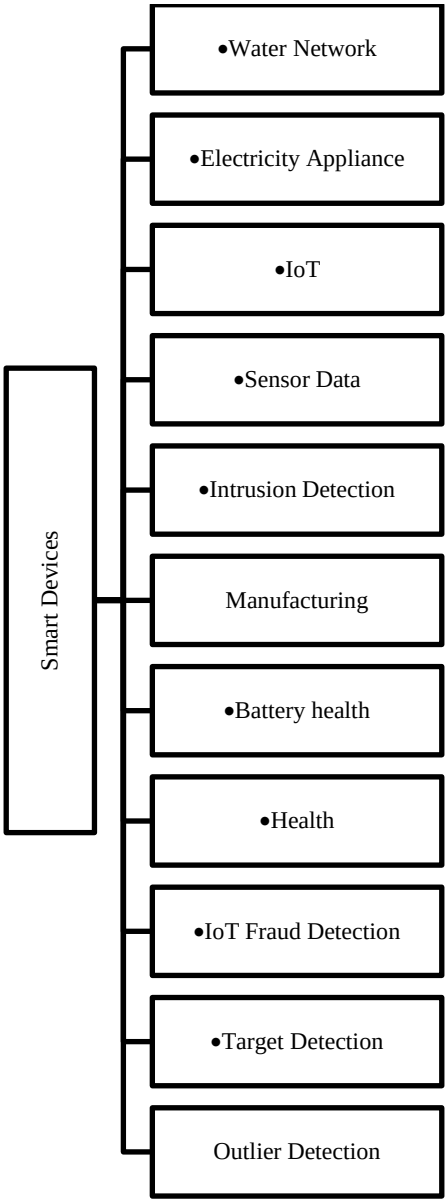


Fig. 18. Classification of Smart Devices Aspect.

TABLE IV. SUMMARY OF PREVIOUS RESEARCH USING ENSEMBLE METHOD IN SMART DEVICES ASPECT

Ref.	Year	Blockchain Application	Application	Ensemble method applied	Model/Based learners	Tools/Dataset
[61]	2021	Blockchain simulator	IoT	Boosting	XGBoost	Bitcoin-transaction, blockchain.info, Bitcoins
[105]	2022	Blockchain-based	IoT	Boosting	Adaboost, Random Forest, DT, NB, KNN	IoTID20

V. Discussion

As demonstrated in Tables I to IV, several studies have been conducted and published since the creation and application of Machine Learning (ML) algorithms in blockchain networks. In this investigation, the researchers' implementation of the ensemble method has demonstrated an improvement pattern. The ensemble strategy is based on combining multiple individual models to generate a model with superior performance compared to a poor classifier. As a result, researchers are continually on the lookout for procedures or processes that provide better results over time than present approaches. Consequently, the strategy of merging multiple ensemble algorithms can give superior results compared to the use of individual ensemble algorithms. Combining stacking and boosting (stacking and boosting) can improve performance, for instance.

According to Fig. 19, 51 percent of the research articles analyzed used the bagging technique, and this technique was used the most in the selected research. Besides, 27 percent utilized the boosting method, while 7 percent applied both the bagging and boosting procedures. In comparison, 5 percent of research articles employed both boosting and stacking. Furthermore, 3 percent employed the stacking and averaging strategy. Lastly, 2 percent of the research studies incorporated both (bagging and voting) and both (bagging and averaging) (bagging and averaging).

According to Fig. 20, we exhibited 17 distinct ML models that academicians have implemented, with the most usually employed being Random Forest (RF) (27 research articles) (27 research papers). On the other side, seven research publications utilized the Extreme Gradient Boosting (XGBoost) model, while four research studies applied Adaptive Boosting (AdaBoost) and Support Vector Machines (SVM) models. In contrast, three of the study articles employed Decision Tree (DT), Naïve Bayes (NB) and K-Nearest Neighbour (KNN).

Analyzing ensemble learning research in cybercrime, security, smart devices, and information processing employing an ensemble approach with distinct techniques (e.g., voting, averaging, stacking, bagging and boosting) for anomaly detection is in blockchain networks. Moreover, we found research in the cybercrime aspect (16 research articles) as the most popular for anomaly identification in the blockchain network. On the other hand, five research publications focused on security aspects, while three research papers focused on information processing. Furthermore, one study paper was applied to the smart device's aspect.

Fig. 21 indicates the fast-increasing tendency of adopting bagging methods in the last four years (from 2017 to 2020) and shows a declining trend in 2021. On the other hand, the research publications utilizing the boosting method show growth from 2017 to 2021. Apart from that, 31 distinct datasets utilized in the experiments of connected papers were found. As depicted in Fig. 22, most experiments utilize real-time datasets retrieved using the Etherscan Application Programme Interface (API).

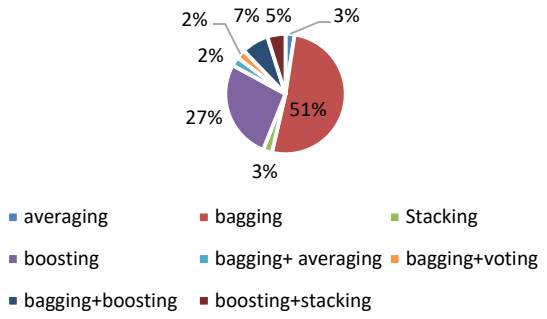


Fig. 19. Percentage of Ensemble Method.

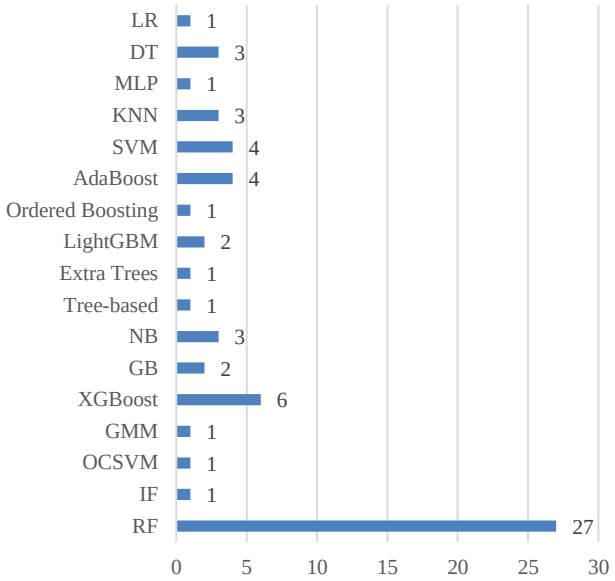


Fig. 20. Frequency of Ensemble Model Base Learner.

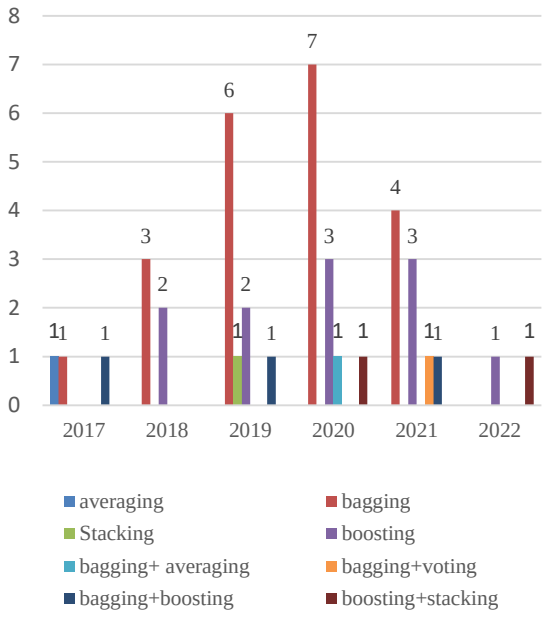


Fig. 21. Anomaly Detection using Ensemble Method Iteration Per Year.

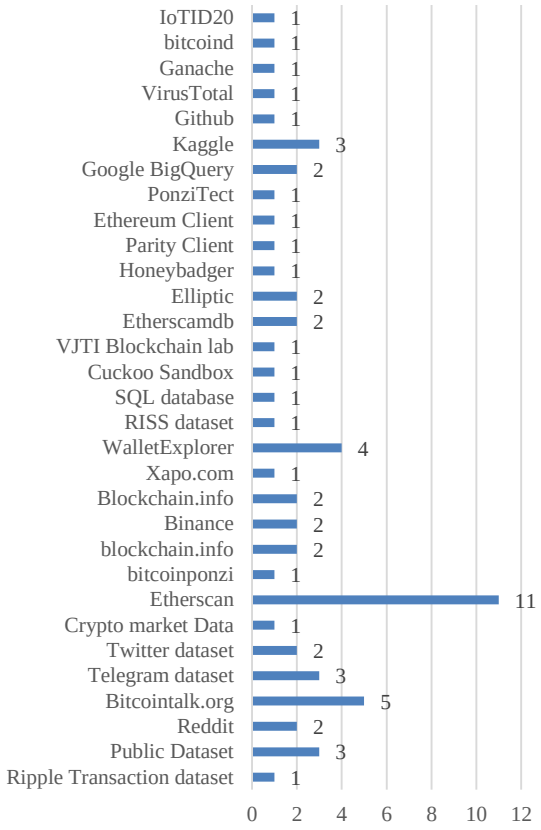


Fig. 22. Utilized Database and Tools in Collected Research Articles.

There are also some prospective challenges in this domain. In addition to analyzing prior studies, several upcoming studies can be highlighted and enhanced. Among these are studies that do not employ feature selection, which has been demonstrated in several prior studies to increase the performance of outcomes. In addition, the majority of studies utilize obsolete data sets. Therefore, it is recommended that researchers regularly update data. This is because scams and cyber assaults contain crucial data in datasets that must be analyzed to develop better trials. This is supported by [106], who concluded that outdated data usage contributed to the efficacy of drop-in attack detection. Furthermore, the authors in [107] concur that researchers should utilize current databases for their studies.

Exploration of new technologies like ML Designer and AutoML affords researchers the option to undertake research. In the study, adapting the strategy of applying feature selection also yielded positive results. This research [74] compared the detection of anomalies using feature selection against those without feature selection. Using synthetic data sources is another way that can aid in the production of more precise research. For example, this strategy was utilized by [108] in employing synthetic credit card data to detect credit card fraud. Additionally, the authors [99] utilized artificial data to imitate network assault activities. Researchers should also look into techniques to automate various preprocessing stages [109], as well as expand and enlarge datasets [110]. In addition, more dedicated preprocessing steps should be adopted for more specific challenges to improve the result of the Ssoft-TeC and give a more appropriate based learner for the co-training scheme [111].

VI. Conclusion

This paper examines the understanding of Blockchain Technology (BT), Blockchain and Machine Learning (ML) integration. It examines previous research on the usage of ensemble approaches as a means of anomaly identification. This investigation demonstrates that assembling strategies can enhance performance and results. The merging of numerous weak models facilitates their unification, resulting in the creation of stronger models. Nevertheless, a mix of ensemble techniques (such as stacking and bagging) can also generate more accurate findings, as demonstrated by several earlier researches.

As demonstrated in Tables I to IV, bagging and boosting are two approaches utilized regularly in the studies over these five years (2017–2020). Nonetheless, we can note that these two strategies are delivering the greatest outcomes largely among research released in 2019 and 2020. In the past two years, we also observed a new trend toward the use of the boosting method. Moreover, from the model employed in the ensemble learning approach, Random Forest (RF) dominated from 2017 to 2020. In 2021, this model declined, whereas Extreme Gradient Boosting (XGBoost) exhibited a growing tendency from 2017 to 2021.

Acknowledgment

This research was conducted to fulfil the requirements for a PhD and with the support of RMIC (UniSZA).

References

[1] S. Nakamoto, “Bitcoin: a peer-to-peer electronic cash system, October 2008,” Cited on, 2008.

[2] Z. Li, R. Y. Zhong, Z. G. Tian, H. N. Dai, A. V. Barenji, and G. Q. Huang, “Industrial Blockchain: A state-of-the-art Survey,” Robotics and Computer-Integrated Manufacturing. 2021, doi: 10.1016/j.rcim.2021.102124.

[3] Y. Xiao, N. Zhang, W. Lou, and Y. T. Hou, “A Survey of Distributed Consensus Protocols for Blockchain Networks,” IEEE Commun. Surv. Tutorials, 2020, doi: 10.1109/COMST.2020.2969706.

[4] S. Wang, L. Ouyang, Y. Yuan, X. Ni, X. Han, and F. Y. Wang, “Blockchain-Enabled Smart Contracts: Architecture, Applications, and Future Trends,” IEEE Trans. Syst. Man, Cybern. Syst., 2019, doi: 10.1109/TSMC.2019.2895123.

[5] M. Rahouti, K. Xiong, and N. Ghani, “Bitcoin Concepts, Threats, and Machine-Learning Security Solutions,” IEEE Access, 2018, doi: 10.1109/ACCESS.2018.2874539.

[6] M. Conti, K. E. Sandeep, C. Lal, and S. Ruj, “A survey on security and privacy issues of bitcoin,” IEEE Commun. Surv. Tutorials, 2018, doi: 10.1109/COMST.2018.2842460.

[7] T. H. A. Musa and A. Bouras, “Anomaly Detection: A Survey,” 2022, doi: 10.1007/978-981-16-2102-4_36.

[8] A. H. Mohsin et al., “Blockchain authentication of network applications: Taxonomy, classification, capabilities, open challenges, motivations, recommendations and future directions,” Computer Standards and Interfaces. 2019, doi: 10.1016/j.csi.2018.12.002.

[9] D. Preuveneers, V. Rimmer, I. Tsingenopoulos, J. Spooren, W. Joosen, and E. Ilie-Zudor, “Chained anomaly detection models for federated learning: An intrusion detection case study,” Appl. Sci., vol. 8, no. 12, pp. 1–21, 2018, doi: 10.3390/app8122663.

[10] W. Sun and B. Trevor, "A stacking ensemble learning framework for annual river ice breakup dates," J. Hydrol., 2018, doi: 10.1016/j.jhydrol.2018.04.008.

[11] S. Haber and W. S. Stornetta, "How to time-stamp a digital document. In Conference on the Theory and Application of Cryptography," 1990.

[12] G. Becker, "Merkle Signature Schemes , Merkle Trees and Their Cryptanalysis."

[13] N. Bozic, G. Pujolle, and S. Secci, "A tutorial on blockchain and applications to secure network control-planes," 2017, doi: 10.1109/SCNS.2016.7870552.

[14] M. Vukolić, "The quest for scalable blockchain fabric: Proof-of-work vs. BFT replication," 2016, doi: 10.1007/978-3-319-39028-4_9.

[15] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends," 2017, doi: 10.1109/BigDataCongress.2017.85.

[16] S. Thakur and V. Kulkarni, "Blockchain and Its Applications – A Detailed Survey," Int. J. Comput. Appl., 2017, doi: 10.5120/ijca2017915994.

[17] M. Vukolić, "Rethinking permissioned blockchains," 2017, doi: 10.1145/3055518.3055526.

[18] Z. Li, J. Kang, R. Yu, D. Ye, Q. Deng, and Y. Zhang, "Consortium blockchain for secure energy trading in industrial internet of things," IEEE Trans. Ind. Informatics, 2018, doi: 10.1109/TII.2017.2786307.

[19] V. Buterin, "A next-generation smart contract and decentralized application platform," Etherum, 2014.

[20] Y. Yuan and F. Y. Wang, "Blockchain and Cryptocurrencies: Model, Techniques, and Applications," IEEE Trans. Syst. Man, Cybern. Syst., 2018, doi: 10.1109/TSMC.2018.2854904.

[21] N. Modiri, "The ISO Reference Model Entities," IEEE Netw., 1991, doi: 10.1109/65.93182.

[22] A. Bogner, M. Chanson, and A. Meeuw, "A decentralised sharing app running a smart contract on the ethereum blockchain," 2016, doi: 10.1145/2991561.2998465.

[23] W. Y. M. M. Thin, N. Dong, G. Bai, and J. S. Dong, "Formal analysis of a proof-of-stake blockchain," 2018, doi: 10.1109/ICECCS2018.2018.00031.

[24] L. M. Bach, B. Mihaljevic, and M. Zagar, "Comparative analysis of blockchain consensus algorithms," 2018, doi: 10.23919/MIPRO.2018.8400278.

[25] M. Castro and B. Liskov, "Practical Byzantine Fault Tolerance and Proactive Recovery," ACM Trans. Comput. Syst., 2002, doi: 10.1145/571637.571640.

[26] F. Tschorsch and B. Scheuermann, "Bitcoin and beyond: A technical survey on decentralized digital currencies," IEEE Commun. Surv. Tutorials, 2016, doi: 10.1109/COMST.2016.2535718.

[27] Buterin and Vitalik, "Ethereum White Paper: A Next Generation Smart Contract & Decentralized Application Platform," Etherum, 2014.

[28] G. Wood, "ETHEREUM: A SECURE DECENTRALISED GENERALISED TRANSACTION LEDGER - BYZANTIUM VERSION 14c313b," Ethereum Proj. Yellow Pap., 2018.

[29] Nick Szabo, "Nick Szabo. The idea of smart contracts.pdf." 1997.

[30] "Winning the Netflix Prize: A Summary," 2011. http://blog.echen.me/2011/10/24/winning-the-netflix-prize-a-summary/ (accessed Apr. 18, 2022).

[31] A. Niculescu-mizil, C. Perlich, G. Swirszcz, and V. Sind-, "Winning the KDD Cup Orange Challenge with Ensemble Selection," pp. 23–34, 2009.

[32] M. Zounemat-Kermani, D. Stephan, M. Barjenbruch, and R. Hinkelmann, "Ensemble data mining modeling in corrosion of concrete sewer: A comparative study of network-based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models," Adv. Eng. Informatics, 2020, doi: 10.1016/j.aei.2019.101030.

[33] B. Baesens, V. Van Vlasselaer, and W. Verbeke, "Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection," J. Chem. Inf. Model., 2015.

[34] A. Chiang and Y. R. Yeh, "Anomaly detection ensembles: in defense of the average," 2016, doi: 10.1109/WI-IAT.2015.260.

[35] I. Alarab, S. Prakoonwit, and M. I. Nacer, "Comparative Analysis Using Supervised Learning Methods for Anti-Money Laundering in Bitcoin."

[36] C. C. Aggarwal, Data classification: Algorithms and applications. 2014.

[37] T. Journal, "An ensemble based approach for effective intrusion detection using majority voting," doi: 10.12928/TELKOMNIKA.v19i2.18325.

[38] K. Xu, M. Xia, X. Mu, Y. Wang, and N. Cao, "EnsembleLens : Ensemble-based Visual Exploration of Anomaly Detection Algorithms with Multidimensional Data," vol. 25, no. 1, pp. 109–119, 2019.

[39] A. Chiang, E. David, Y. Lee, G. Leshem, and Y. Yeh, "A study on anomaly detection ensembles," J. Appl. Log., vol. 21, pp. 1–13, 2017, doi: 10.1016/j.jal.2016.12.002.

[40] D. H. Wolpert, "Stacked Generalization This work was performed under the auspices of the Department of Energy . LA-UR-90-3460," vol. 6080, no. December, 2018, doi: 10.1016/S0893-6080(05)80023-1.

[41] W. Sun and Z. Li, "Hourly PM2.5 concentration forecasting based on feature extraction and stacking-driven ensemble model for the winter of the Beijing-Tianjin-Hebei area," Atmos. Pollut. Res., 2020, doi: 10.1016/j.apr.2020.02.022.

[42] M. Zounemat-kermani, O. Batelaan, M. Fadaee, and R. Hinkelmann, "Ensemble machine learning paradigms in hydrology : A review," vol. 598, no. December 2020, 2021.

[43] L. Breiman, "Bagging predictors," Mach. Learn., 1996, doi: 10.1007/bf00058655.

[44] Y. Wu, Y. Ke, Z. Chen, S. Liang, H. Zhao, and H. Hong, "Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping," Catena, 2020, doi: 10.1016/j.catena.2019.104396.

[45] R. M. Adnan, Z. Liang, S. Heddam, M. Zounemat-Kermani, O. Kisi, and B. Li, "Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs," J. Hydrol., 2020, doi: 10.1016/j.jhydrol.2019.124371.

[46] F. T. Liu and K. M. Ting, "Isolation Forest," 2008, doi: 10.1109/ICDM.2008.17.

[47] Paul, "Bagging, Boosting, Stacking and Cascading Classifiers in Machine Learning using SKLEARN and MLEXTEND," 2018. https://www.mendeley.com/search/?page=1&query=Bagging%2C Boosting%2C Stacking and Cascading Classifiers in Machine Learning using SKLEARN and MLEXTEND&sortBy=relevance (accessed Apr. 14, 2022).

[48] R. E. Schapire, "The Boosting Approach to Machine Learning: An Overview BT - Nonlinear Estimation and Classification," Nonlinear Estim. Classif., 2003.

[49] E. Alfaro, M. Gáamez, and N. García, "Adabag: An R package for classification with boosting and bagging," J. Stat. Softw., 2013, doi: 10.18637/jss.v054.i02.

[50] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," J. Comput. Syst. Sci., 1997, doi: 10.1006/jcss.1997.1504.

[51] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[52] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., 2001, doi: 10.1214/aos/1013203451.

[53] M. Ostapowicz and K. Żbikowski, "Detecting Fraudulent Accounts on Blockchain: A Supervised Approach," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11881 LNCS, pp. 18–31, 2019, doi: 10.1007/978-3-030-34223-4_2.

[54] G. Ke et al., "LightGBM : A Highly Efficient Gradient Boosting Decision Tree," no. Nips, pp. 1–9, 2017.

[55] CoinMarketCap, "Today's price, charts, and info | Crypto-Currency Market Capitalizations," coinmarketcap.com, 2022..

[56] "Ethereum Transactions Per Day," 2022. https://ycharts.com/indicators/ethereum_transactions_per_day (accessed May 11, 2022).

[57] N. Kumar, A. Singh, A. Handa, and S. K. Shukla, "Detecting Malicious Accounts on the Ethereum Blockchain with Supervised Learning," 2020, doi: 10.1007/978-3-030-49785-9_7.

[58] P. Monamo, V. Marivate, and B. Twala, "Unsupervised learning for robust Bitcoin fraud detection," 2016 Inf. Secur. South Africa - Proc. 2016 ISSA Conf., pp. 129–134, 2016, doi: 10.1109/ISSA.2016.7802939.

[59] R. F. Ibrahim, A. M. Elian, and M. Ababneh, "Illicit Account Detection in the Ethereum Blockchain Using Machine Learning," 2021 Int. Conf. Inf. Technol. ICIT 2021 - Proc., pp. 488–493, 2021, doi: 10.1109/ICIT52682.2021.9491653.

[60] P. Nerurkar, Y. Busnel, R. Ludinard, K. Shah, S. Bhirud, and D. Patel, "Detecting Illicit Entities in Bitcoin using Supervised Learning of Ensemble Decision Trees," ACM Int. Conf. Proceeding Ser., pp. 25–30, 2020, doi: 10.1145/3418981.3418984.

[61] D. K. Gouda, S. Jolly, and K. Kapoor, "Design and Validation of BlockEval , A Blockchain Simulator," vol. 2061, pp. 281–289.

[62] H. Wen, J. Fang, J. Wu, and Z. Zheng, "Transaction-based Hidden Strategies Against General Phishing Detection Framework on Ethereum," 2021.

[63] M. Shen, X. Tang, L. Zhu, X. Du, and M. Guizani, "Privacy-Preserving Support Vector Machine Training over Blockchain-Based Encrypted IoT Data in Smart Cities," IEEE Internet Things J., vol. 6, no. 5, pp. 7702–7712, 2019, doi: 10.1109/JIOT.2019.2901840.

[64] Y. J. Lin, P. W. Wu, C. H. Hsu, I. P. Tu, and S. W. Liao, "An Evaluation of Bitcoin Address Classification based on Transaction History Summarization," ICBC 2019 - IEEE Int. Conf. Blockchain Cryptocurrency, pp. 302–310, 2019, doi: 10.1109/BLOC.2019.8751410.

[65] M. Li, K. Zhang, J. Liu, H. Gong, and Z. Zhang, "Blockchain-based anomaly detection of electricity consumption in smart grids," Pattern Recognit. Lett., vol. 138, pp. 476–482, 2020, doi: 10.1016/j.patrec.2020.07.020.

[66] Q. Ngo, H. Nguyen, H. Tran, and D. Nguyen, "IoT Botnet detection based on the integration of static and dynamic vector features," pp. 540–545, 2020.

[67] P. Barlet-ros, "Detecting cryptocurrency miners with NetFlow / IPFIX network measurements," 2019.

[68] X. Liu, F. Jiang, and R. Zhang, "A New Social User Anomaly Behavior Detection System Based on Blockchain and Smart Contract," 2020 IEEE Int. Conf. Networking, Sens. Control. ICNSC 2020, 2020, doi: 10.1109/ICNSC48988.2020.9238118.

[69] J. Wu et al., "Who Are the Phishers? Phishing Scam Detection on Ethereum via Network Embedding," IEEE Trans. Syst. Man, Cybern. Syst., vol. 52, no. 2, pp. 1156–1166, 2022, doi: 10.1109/TSMC.2020.3016821.

[70] M. S. Bhargavi, S. M. Katti, M. Shilpa, V. P. Kulkarni, and S. Prasad, "Transactional Data Analytics for Inferring Behavioural Traits in Ethereum Blockchain Network," Proc. - 2020 IEEE 16th Int. Conf. Intell. Comput. Commun. Process. ICCP 2020, pp. 485–490, 2020, doi: 10.1109/ICCP51029.2020.9266176.

[71] S. Iyer, S. Thakur, M. DIxit, R. Katkam, A. Agrawal, and F. Kazi, "Blockchain and Anomaly Detection based Monitoring System for Enforcing Wastewater Reuse," 2019 10th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2019, 2019, doi: 10.1109/ICCCNT45670.2019.8944586.

[72] M. Keshk, B. Turnbull, N. Moustafa, D. Vatsalan, K. R. Choo, and S. Member, "A Privacy-Preserving-Framework-Based Blockchain and Deep Learning for Protecting Smart Power Networks," vol. 16, no. 8, pp. 5110–5118, 2020.

[73] F. Poursafaei, G. B. Hamad, and Z. Zilic, "Detecting Malicious Ethereum Entities via Application of Machine Learning Classification," 2020 2nd Conf. Blockchain Res. Appl. Innov. Networks Serv. BRAINS 2020, pp. 120–127, 2020, doi: 10.1109/BRAINS49436.2020.9223304.

[74] C. Jatoth, R. Jain, U. Fiore, and S. Chatharasupalli, "Improved Classification of Blockchain Transactions Using Feature Engineering and Ensemble Learning," Futur. Internet, vol. 14, no. 1, pp. 1–13, 2022, doi: 10.3390/fi14010016.

[75] A. Ahmed, M. Mubashir, K. Mehboob, J. Arshad, and F. Ahmad, "A blockchain-based decentralized machine learning framework for collaborative intrusion detection within UAVs," Comput. Networks, vol. 196, no. December 2020, p. 108217, 2021, doi: 10.1016/j.comnet.2021.108217.

[76] C. M. M. Reep-van den Bergh and M. Junger, "Victims of cybercrime in Europe: a review of victim surveys," Crime Sci., 2018, doi: 10.1186/s40163-018-0079-3.

[77] S. H. Kok, A. Abdullah, and N. Z. Jhanjhi, "Prevention of Crypto-Ransomware Using a Pre-Encryption Detection Algorithm," pp. 1–15, 2019.

[78] M. Bhowmik, T. Sai Siri Chandana, and B. Rudra, "Comparative Study of Machine Learning Algorithms for Fraud Detection in Blockchain," Proc. - 5th Int. Conf. Comput. Methodol. Commun. ICCMC 2021, no. Iccmc, pp. 539–541, 2021, doi: 10.1109/ICCMC51019.2021.9418470.

[79] P. M. Monamo, V. Marivate, and B. Twala, "A Multifaceted Approach to Bitcoin Fraud Detection: Global and Local Outliers," no. December, pp. 188–194, 2017, doi: 10.1109/icmla.2016.0039.

[80] R. D. Camino, R. State, L. Montero, and P. Valtchev, "Finding suspicious activities in financial transactions and distributed ledgers," IEEE Int. Conf. Data Min. Work. ICDMW, vol. 2017-Novem, pp. 787–796, 2017, doi: 10.1109/ICDMW.2017.109.

[81] K. Toyoda, T. Ohtsuki, and P. T. Mathiopoulos, "Identification of High Yielding Investment Programs in Bitcoin via Transactions Pattern Analysis," 2017.

[82] M. Bartoletti, B. Pes, and S. Serusi, "Data mining for detecting bitcoin ponzi schemes," Proc. - 2018 Crypto Val. Conf. Blockchain Technol. CVCBT 2018, pp. 75–84, 2018, doi: 10.1109/CVCBT.2018.00014.

[83] M. Mirtaheri, F. Morstatter, and G. Ver Steeg, "Identifying and Analyzing Cryptocurrency Manipulations in Social Media," 2018.

[84] W. Chen, Z. Zheng, J. Cui, E. Ngai, P. Zheng, and Y. Zhou, "Detecting ponzi schemes on ethereum: Towards healthier blockchain technology," Web Conf. 2018 - Proc. World Wide Web Conf. WWW 2018, pp. 1409–1418, 2018, doi: 10.1145/3178876.3186046.

[85] A. Harlev, H. S. Yin, and K. C. Langenheldt, "Breaking Bad : De-Anonymising Entity Types on the Bitcoin Blockchain Using Supervised Machine Learning," vol. 9, pp. 3497–3506, 2018.

[86] H. Baek, J. Oh, C. Y. Kim, and K. Lee, "A Model for Detecting Cryptocurrency Transactions with Discernible Purpose," Int. Conf. Ubiquitous Futur. Networks, ICUFN, vol. 2019-July, pp. 713–717, 2019, doi: 10.1109/ICUFN.2019.8806126.

[87] W. Chen, Z. Zheng, E. C. H. Ngai, P. Zheng, and Y. Zhou, "Exploiting Blockchain Data to Detect Smart Ponzi Schemes on Ethereum," IEEE Access, vol. 7, pp. 37575–37586, 2019, doi: 10.1109/ACCESS.2019.2905769.

[88] K. Toyoda, P. T. Mathiopoulos, and T. Ohtsuki, "A Novel Methodology for HYIP Operators ' Bitcoin Addresses Identification," IEEE Access, vol. 7, pp. 74835–74848, 2019, doi: 10.1109/ACCESS.2019.2921087.

[89] S. Farrugia, J. Ellul, and G. Azzopardi, "Detection of illicit accounts over the Ethereum blockchain," Expert Syst. Appl., vol. 150, no. February 2019, p. 113318, 2020, doi: 10.1016/j.eswa.2020.113318.

[90] J. Lorenz, M. I. Silva, D. Aparício, J. T. Ascensão, and P. Bizarro, "Machine learning methods to detect money laundering in the bitcoin blockchain in the presence of label scarcity," ICAIF 2020 - 1st ACM Int. Conf. AI Financ., 2020, doi: 10.1145/3383455.3422549.

[91] K. Lašas, G. Kasputyte, R. Užupyte, and T. Krilavičius, "Fraudulent behaviour identification in ethereum blockchain," CEUR Workshop Proc., vol. 2698, 2020.

[92] W. Chen, Z. Chen, and Y. Lu, "Honeypot Contract Risk Warning on Ethereum Smart Contracts," pp. 1–8, 2020, doi: 10.1109/JCC49151.2020.00009.

[93] S. Fan, S. Fu, H. Xu, and C. Zhu, "Expose Your Mask : Smart Ponzi Schemes Detection on Blockchain," no. September 2014, 2020.

[94] D. Boughaci, "Enhancing the security of financial transactions in Blockchain by using machine learning techniques : towards a sophisticated security tool for banking and finance," pp. 110–115, 2020, doi: 10.1109/SMART-TECH49988.2020.00038.

[95] M. La Morgia, A. Mei, F. Sassi, and J. Stefa, "Pump and Dumps in the Bitcoin Era : Real Time Detection of Cryptocurrency Market Manipulations," 2020.

[96] B. Chen, F. Wei, and C. Gu, "Bitcoin Theft Detection Based on Supervised Machine Learning Algorithms," Secur. Commun. Networks, vol. 2021, no. August 2016, 2021, doi: 10.1155/2021/6643763.

[97] S. Al-e, M. Anbar, Y. Sanjalawe, and S. Manickam, A Labeled Transactions-Based Dataset on the Ethereum Network A Labeled Transactions-Based Dataset on the Ethereum Network, no. July. Springer Singapore, 2021.

[98] W. Wang, J. Song, G. Xu, Y. Li, H. Wang, and C. Su, "ContractWard : Automated Vulnerability Detection Models for Ethereum Smart Contracts," no. January, 2020, doi: 10.1109/TNSE.2020.2968505.

[99] J. Eduardo A. Sousa et al., "Fighting Under-price DoS Attack in Ethereum with Machine Learning Techniques," Perform. Eval. Rev., vol. 48, no. 4, pp. 24–27, 2021, doi: 10.1145/3466826.3466835.

[100] S. Dashevskyi, Y. Zhauniarovich, O. Gadyatskaya, A. Pilgun, and H. Ouhssain, "Dissecting Android Cryptocurrency Miners," 2020, doi: 10.1145/3374664.3375724.

[101] R. Agarwal, S. Barve, and S. K. Shukla, "Detecting malicious accounts in permissionless blockchains using temporal graph properties," Appl. Netw. Sci., vol. 6, no. 1, 2021, doi: 10.1007/s41109-020-00338-3.

[102] F. Zola, J. L. Bruse, M. Eguimendia, M. Galar, and R. O. Urrutia, "applied sciences Bitcoin and Cybersecurity : Temporal Dissection of Blockchain Data to Unveil Changes in Entity Behavioral Patterns," 2019, doi: 10.3390/app9235003.

[103] K. Kanemura, "Identification of Darknet Markets ' Bitcoin Addresses by Voting Per-address Classification Results," pp. 154–158, 2019.

[104] S. Ranshous et al., "Exchange pattern mining in the bitcoin transaction directed hypergraph," 2017, doi: 10.1007/978-3-319-70278-0_16.

[105] R. Shahin and K. E. Sabri, "A Secure IoT Framework Based on Blockchain and Machine Learning," vol. 1, no. 1, 2022.

[106] S. R. Khonde and V. Ulagamuthalvi, "Blockchain : Secured Solution for Signature Transfer in Distributed Intrusion Detection System," 2022, doi: 10.32604/csse.2022.017130.

[107] "Machine_Learning_for_Anomaly_Detection_A_Systematic_Review.pdf.crdownload." .

[108] E. Ileberi, Y. Sun, and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," IEEE Access, vol. 9, pp. 165286–165294, 2021, doi: 10.1109/ACCESS.2021.3134330.

[109] H. Sun Yin and R. Vatrapu, "A first estimation of the proportion of cybercriminal entities in the bitcoin ecosystem using supervised machine learning," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018-Janua, no. March, pp. 3690–3699, 2017, doi: 10.1109/BigData.2017.8258365.

[110] P. Momeni and Y. Wang, "Machine Learning Model for Smart Contracts Security Analysis," 2019.

[111] P. Pintelas and I. E. Livieris, Ensemble Algorithms and Their Applications. 2020.

[112] M. K. Awang, M. Makhtar, N. Udin, and N. F. Mansor, "Improving Customer Churn Classification with Ensemble Stacking Method," vol. 12, no. 11, 2021.

[113] R. Rosly, M. Makhtar, M. K. Awang, H. Hassan, A. Nazari, and M. Rose, "Deep Multi-Classifier Learning for Medical Data Sets," pp. 1–7, 2020.

[114] M. Makhtar, L. Yang, D. Neagu, and M. Ridley, "Optimisation of classifier ensemble for predictive toxicology applications," Proc. - 2012 14th Int. Conf. Model. Simulation, UKSim 2012, pp. 236–241, 2012, doi: 10.1109/UKSim.2012.41.

A Novel Hybrid Sentiment Analysis Classification Approach for Mobile Applications Arabic Slang Reviews

Rabab Emad Saudy[1]
Department of Information Systems Technology
Faculty of Graduated Studies for Statistical Research
Cairo University, Cairo, Egypt

Alaa El Din M. El-Ghazaly[2]
Department of Computer and Information Sciences
Sadat Academy for Management Sciences, Cairo, Egypt

Eman S. Nasr[3]
Independent Researcher
Cairo, Egypt

Mervat H. Gheith[4]
Department of Computer Science
Faculty of Graduated Studies for Statistical Research
Cairo University, Cairo, Egypt

***Abstract*—Arabic language incurs from the shortage of accessible huge datasets for Sentiment Analysis (SA), Machine Learning (ML), and Deep Learning (DL) applications. In this paper, we present MASR, a simple Mobile Applications Arabic Slang Reviews dataset for SA, ML, and DL applications which comprises of 2469 Egyptian Mobile Apps reviews, and help app developers meet user requirements evolution. Our methodology consists of six phases. We collect mobile apps reviews dataset, then apply preprocessing steps, in addition perform SA tasks. To evaluate MASR datasets, first we apply ML classification techniques: K-Nearest Neighbors (K-NN), Support vector machine (SVM), Logistic Regression (LR), and Random Forest (RF), and DL classification technique: Multi-layer Perceptron Neural Network (MLP-NN). From the examination for pervious classification techniques, we adopted a hybrid classification approach combined from the top two ML classifier accuracy results (LR, RF), and DL classifier (MLP-NN). The findings prove the adequacy of a hybrid supervised classification approach for MASR datasets.**

Keywords*—Arabic sentiment analysis; mobile application; hybrid classification model; hybrid supervised classification approach; Google play store; random forest; logistic regression; neural network; multi-layer perceptron neural network; machine learning; deep learning

I. INTRODUCTION

Mobile app stores supply an amazingly wealthy source of information on app specification, characteristics, and utilize, and analyzing these information supplies knowledge and a more profound comprehension of the idea of apps. However, manual analysis of this tremendous measure of information on mobile apps is anything but a basic and clear task; it is expensive as far as human effort and time [1]. There are different mobile app stores, for example, Google, and Apple app store, and others that include free and paid mobile apps [2].

Mobile app classification phase is classified based on a significant category or class. In case users want to investigate and discover an app reasonable for their requirements, it is more helpful to have a special predefined classification scheme by which all apps are classified [3].

Being a significant provenance of data for organizations, the requirement to produce exact SA is a significant issue. Most sentiments accumulated from Arabic resources like social media is in colloquial Arabic, as the utilization of Modern Standard Arabic (MSA) in online is uncommon [4].

A few researches have been directed to analyze English mobile apps [5] [6] [7] [8] [9] [10]. In addition, according to the literature review, few researches have analyzed Islamic Arabic mobile apps and Saudi governmental services mobile apps [1] [11] [12]. However, no previous study has constructed, classified or analyzed Egyptian Dialect Arabic (DA) mobile apps reviews dataset.

The contributions in this research can be summed up as follows:

1) Introduce present MASR, simple Mobile Applications Arabic Slang Reviews of Egyptian reviews dataset for SA, ML and DL applications.

2) Investigate the structure, properties of the dataset, and perform tests on selected attributes for sentiment polarity classification.

3) Apply a various supervised ML, DL classifiers to the simple MASR that we gathered.

4) Adopted a hybrid supervised sentiment analysis classification approach including heterogenous approaches: Machine Learning (ML) approach such as: Logistic Regression (LR), and Random Forest (RF), and Deep Learning (DL) approach: Multi-layer Perceptron Neural Network (MLP-NN) classifiers to enhance the performance models of predicting MASR datasets and accuracy.

5) Compare our proposed model approach performance with various ML, and DL models.

The rest of the paper is organized as follows: Section II presents the literature review. Section III presents the six

phases of our proposed hybrid classification approach methodology. Section IV presents experimental results and discussion. Section V presents conclusion and Section VI presents future works.

II. Literature Review

Slight endeavors have been made to anatomize mobile apps reviews to handle mobile apps requirements evolution, advancement information and significant software. Related previous studies handle many aspects in mining mobile apps reviews for different sentiment analysis purposes such as building lexicons, classifying non-functional requirements, classify buggy apps, recognizing high-rated apps, and hybrid system to find the most similar word in lexicon for Egyptian Arabic tweets.

1) Arabic sentiment analysis tasks: El-Beltagy et al. [13] build a sentimental Egyptian Dialect lexicon. Their tests showed that their proposed methodology gave improved results with regards to twitter even with the poor utilized resources.

Fu et al. [14] dealt with an enormous user reviews dataset including about 13 million mobile apps reviews from google play store. The creators proposed a WisCom framework to recognize the motivations behind why clients dislike specific mobile apps.

Gómez et al. [15] construct mobile apps reviews dataset to evolve a framework that identifies conceivably buggy mobile apps by enforcing a linkage in consent patterns and fault related reviews.

Chen et al. [16] presented a SimApp framework for identifying similar apps utilizing machine learning algorithms. SimApp inspects multimodal different data in app stores. They construct numerous kernel functions to degree app similarity. The outcomes exhibit that SimApp is powerful and promising for use in numerous applications, for example, app categorization, search and recommendation.

Tian et al. [5] research the main factors for recognizing high-rated apps by implementing random forest classifier. The test indicates that the main factors are promotional images numbers appeared on the app page, app size, and app version.

Lu et al. [17] suggest an approach to deal with classify mobile apps reviews automatically in light of non-functional requirements. They gathered 11,096 mobile apps reviews from Apple Store and Google Play.

Hameed et al. [11] explore existing Islamic apps accessible on Google Play app store. They handled the issue of the shortfall classification and the mis-categorization of Islamic apps. Therefore, they recommended another categorization for the Islamic apps' dependent on their common features such as download numbers, app ratings, and languages. They gathered proposed 5 distinct classes for the Islamic apps: Zakat, Qibla/Prayer Time, Quran, Hadith, and Supplications.

Abuelenin et al. [18] proposed hybrid system to find the most similar word in lexicon and increase the accuracy of Egyptian Arabic using the cosine similarity algorithm and the Information Science Research Institute Arabic stemmer (ISRI).

Al-Shamani et al. [12] construct Arb-AppsReview dataset for various research domains, such as gender detection, dialect analysis, sentiment analysis.

2) State-of-arts hybrid models: Heikal et al [19] propose a model which applies a hybrid model consists of CNN, and LSTM on ASTD. This model prediction performance is to 65%.

Al-Twairesh et al [20] suggest a model which applies a hybrid model SF+ GE + ASEH on SemEval. This model prediction performance is to 80.36%.

Mohammed et al. [21] propose a model which applies a hybrid model LSTM+Augmented on Arabic tweets. This model prediction performance is to 88.05%.

Furthermore, few previous works suggested a hybrid classification SA model for classify Egyptian Dialect Arabic mobile apps reviews.

III. A Hybrid Sentiment Analysis Classification Approach for Mobile Apps Arabic Slang Reiews (MASR) Methodology

This paper methodology depends on previous qualitative, quantitative and SLR research methodology [22]. It built according to previous observations after analyzing ASA survey, comparative framework [23] and future relationship hypothesis, user satisfaction surveys and case studies. The proposed methodology will be based on applying Natural Processing Language (NLP) and Data Mining (DM) Tools, Methods and Techniques. It depends on the quality of extracted features that express user opinion and its sentiment for Arabic Mobile Apps'. Finally, the main goal for it is to help developers improve and enhance new releases of Mobile Apps to meet rapidly changing in requirements evolution.

This research adopted a hybrid classification model which consist of six phases for collect, analyze and classify sentimental Arabic Dialect mobile apps reviews on google play store, as shown in Fig. 1.

This paper construct six phases for a hybrid classification Model methodology as indicated by Fig. 1; phase 1 MASR collection phase involves how to scrape and gather the dataset from google play store via Appbo[1] scraper tool and describing the dataset characteristics. The second phase involves the implementing of various pre-processing steps which will be applied on MASR dataset. The third phase is implementing feature extraction using Bag of Words (BOW) and Tf-idf. The fourth phase is implementing famous supervised machine learning classification algorithms such as Support Vector Machine (SVM), Naïve Bayes (NB), Linear Regression (LR), Neural Network (NN), and KNN classifier. The fifth phase proposing hybrid classification techniques according to the results of classifiers which accomplish highly accuracy results from the previous phase to enhance MASR accuracy results.

[1] https://appbot.co/

The last phase is to evaluate and compare the classification results utilizing recall, precision and accuracy.

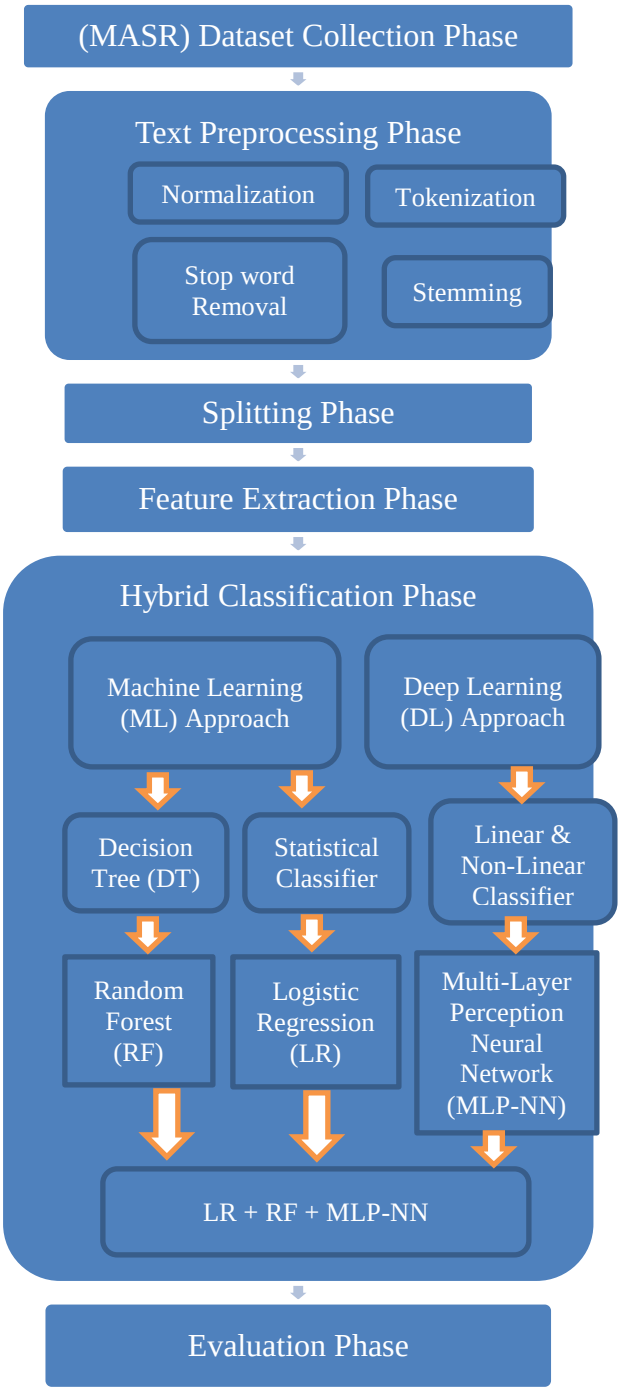


Fig. 1. HSACA-MASR Methodology Phases.

1) First Phase: Mobile Apps Arabic Slang Reviews (MASR) Dataset Collection Phase.

In this research, the Mobile Apps Egyptian DA reviews dataset was extracted using Appbot scraper tool which follows those steps:

- Choose Google Play Store.
- Select 9 various categories of mobile apps as shown in Table I.
- Focus on reviews of Egyptian mobile apps, and another Egyptian reviews for non-Egyptian mobile apps such as: Instagram, as shown in Table I.
- Save extracted attributes and reviews in CSV file: app category, app name, review, rating, and review polarity, as shown in Table II.

TABLE I. APP CATEGORY, APP NAME, APP RATING

	App category	*App name*	*App rating*
1	**Social**	Instagram[2]	4.4
2	**Lifetyle**	ContactCars[3]	4.5
3	**Travel & Locals**	Egypt Air[4]	4
4	**Shopping**	Kazyon[5]	4
		Olx Egypt[6]	4.3
5	**Tools**	Otlob[7]	4
		Shareit[8]	4.1
6	**Medical**	Vezeeta[9]	4.7
7	**Productivity**	Ana Vodafone[10]	4.2
8	**Education**	Aladwaa Education[11]	4
		بنك المعرفة المصرى [12]	4
9	**Maps & Navigation**	Careem[13]	4.2

2) MASR Properties: MASR dataset comprises of 2469 reviews made up of 653 positive, 756 neutral and 1060 negative reviews. A negative review is characterized as a review that has been given a rating of "1" or "2" or "3". A positive review is one where the review has been given a rating of "3" or "4" or "5". At last, Neutral reviews with a rating of "1" or "2" or "3" or "4" or "5". The MASR dataset was made from the gathered data and comprises of the following fundamental attributes as shown in Table II.

3) MASR distribution: MASR dataset covers 2469 mobile apps reviews contributed by various reviewers from 12 mobile apps which covers nine various mobile apps categories such as social, lifestyle, education, maps and navigation, productivity, shopping, travel and tools. The negative reviews comprise 43% of the absolute number of reviews when contrasted with

[2] https://play.google.com/store/apps/details?id=com.instagram.android
[3] https://play.google.com/store/apps/details?id=net.sarmady.contactcarswithtabs
[4] https://play.google.com/store/apps/details?id=com.linkdev.egyptair.app
[5] https://play.google.com/store/apps/details?id=com.inova.kazyon
[6] https://play.google.com/store/apps/details?id=com.olxmena.horizontal
[7] https://play.google.com/store/apps/details?id=com.semicoloneg.otlob
[8] https://play.google.com/store/apps/details?id=com.lenovo.anyshare.gps
[9] https://play.google.com/store/apps/details?id=com.ionicframework.vezeetapatientsmobile694843
[10] https://play.google.com/store/apps/details?id=com.emeint.android.myservices
[11] https://play.google.com/store/apps/details?id=com.nahdetmisr.adwaa
[12] https://play.google.com/store/apps/details?id=banke.elma3regypt
[13] https://play.google.com/store/apps/details?id=com.careem.acma

the 26% of the positive ones. Furthermore, 31% of the reviews are "neutral". As expected, the negative reviews are the greater part class. Fig. 2 presents the classification of ratings for our extracted dataset.

TABLE II. MASR DATASET ATTRIBUTES

Attribute	*Description*
Mobile App Category	Category of mobile app which include various category according to extracted datasets (Social, Lifestyle, Travel & Local, Shopping, Tools, Medical, Productivity, Education, or Maps & Navigation).
Mobile App Name	Name of selected Mobile App.
Review	opinion of reviewer's written in the ED which is mixing between MSA or DA.
Rating	Applies scale from 1 to 5 showing the scope of the reviewer's satisfaction. Positive reviews instead of using the previous scale from 1 to 10.
Review Polarity	Denotes the sentiment of the review with "+1" for a positive review, "−1" for a negative review, and "0" for a neutral review.

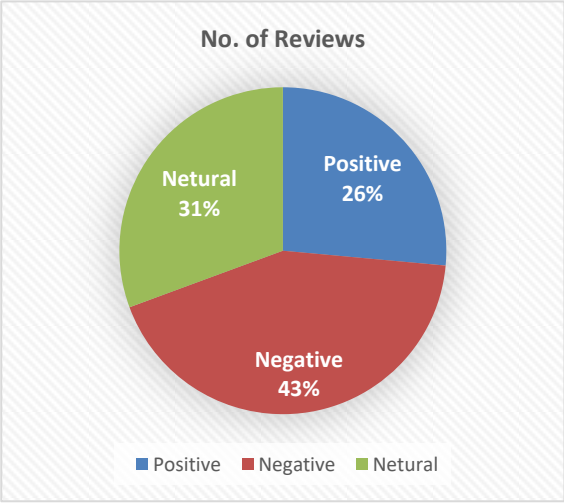


Fig. 2. MASR Dataset Polarity Distribution.

Table III represents samples of MASR datasets which contains app category, app name, review, translated review, rating, and polarity (Negative, Positive, Neutral).

4) Second phase: Text Preprocessing phase: The initial step is to implement text pre-processing so as to evolve the performance of classifiers by changing the text into a format as suitable as possible. To achieve this, many stages are executed; specifically, normalization, tokenization, stop-word removal and stemming.

5) Normalization: This stage includes the accompanying steps: Remove punctuation marks and special characters, remove tatweel kashida symbol ("--"), remove of all diacritics, remove digit numbers (0-9), remove repeated characters, remove all non-Arabic words, replace each final letter (ي) with (ى), replace initial letter alef-hamza (أ، إ،ء،ئ،ؤ، آ) with (ا), and replace each final letter (ة) with (ه).

TABLE III. MASR REVIEWS

App Category	*App Name*	*Review*	*Translated Review*	*Rating*	*Polarity*
Productivity	Ana Vodafone	معرفتش ادفع ببطاقة الائتمان	I don't pay by credit card	3	Negative
Travel & Local	EGYPTAIR	برنامج رائع برجاء اضافة صاله السفر والوصول في حالة الرحلة	Wonderful program, please add the travel and arrival hall in case of flight	4	Positive
Tools	SHAREit	طبيق ممتاااز بس احيانا لما احاول ابعت حاجه لإصدار اقل مني ميبعتش	An excellent application, but sometimes when I try to send something less than me, I need to issue it	5	Neutral

6) Tokenization: For author/s of more than two affiliations: To change the default, adjust the template as follows. By tokenizing, you can appropriately separate text by word or by sentence. This will permit act with smaller sets of text that are still comparatively meaningful regular outgoing of the context of the remainder of the text. In this research, Regexp Nltk14 method applies on MASR dataset. It divides a string into substrings utilizing a standard expression. It can utilize its regexp to look like delimiters instead.

7) Stop word removal: The second stage is to eliminate all stop-words from the reviews. Stop words are characterized as words that don't increase any sentiment value to a review; they are typically the most widely recognized words in a language. They can either be specially made or gained from the web. Unfortunately, there is no clear list accessible and there are slight lists accessible for the Arabic language. This research adjusted Arabic stopword list from many resources in addition to Egyptian stopword list from [24].

8) Stemming: Stemming is a text processing method of decreasing a word to its root. It maps various patterns of the similar word to a public "stem" - for example, the Arabic stemmer maps طفل, اطفال, الاطفال, اطفالكم, فأطفالكم, اطفالهم, والاطفال, والطفولة, وطفل, فاطفالهم, and طفلتان to طفل. In this research, Snowball[15] stemmer applies on MASR dataset.

a) Third phase: Splitting phase: MASR dataset was separated into two sections: training sets, and testing sets. The training sets represent 70% of the datasets, and the testing sets represents 30%. The training sets utilized to train models, while the testing sets utilized to evaluate models.

[14] https://www.nltk.org/_modules/nltk/tokenize/regexp.html
[15] https://git.texta.ee/texta/snowball/-/blob/master/python/testapp.py

b) Fourth phase: Feature extraction phase to estimate classifiers performance, this research utilized various variety of features. Those features can be Bag-of-Words (BOW) with TF-IDF (Term Frequency Inverse Document Frequency).

9) Bag of Words (BOW) [16]: BOW is a process of eliciting features from text for utilize in modeling, such as with ML algorithms. BOW model assigns a corpus with word counts for every document.

*10)*Term Frequency- Inverse Document Frequency (TF-IDF) [17]: Tf-IDF weight is a statistical measure utilized to estimate how significant a word is to a document in a corpus. The significance grows proportionally to the frequency of times a word represents in the document. It is formed by two sections:

$$Tfidf = \log(word, review) * \log \frac{\sum reviews}{\sum freq\ of\ words} \quad (1)$$

a) Fifth phase: Hybrid supervised classification approach phase: This phase performs two subsections: the first issue is applying ML approach which performs five selected ML classifiers which utilized extensively for ASA: Logistic Regression (LR) [25] [26] [27] [28], Naïve Bayes (NB) [29] [30] [31] [32], K-Nearest Neighbors (KNN) [33] [34] [31] [35], Random Forest (RF) [36] [37] [38] and SVM [33] [39] [40] in addition applying DL approach which performs DL classifier Multi-Layer Perceptron Neural Network (MLP-NN) which applied in [36] [37] for ASA.

For the second issue: This research intends to propose a novel Hybrid Supervised Classification Approach to automatically classify and predict the polarity of mobile apps Arabic Slang user reviews. This model mixes various supervised ML, and DL approaches. In ML approach, we suggest various modeling approaches: decision tree approach, and statistical approach. While in DL approach, we suggest linear & non-linear approach. In decision tree approach, we apply RF classifier. In Linear & Non-Linear approach, we apply MLP-NN classifier. In Statistical approach, we apply LR classifier. The reason for selecting those classifiers came after applying various ML classifiers in a previous phase. The results shows that the top classifiers that gain best accuracy for classify or predict MASR datasets are: RF, LR, and MLP-NN. Finally, we propose to apply a hybrid classification model that combines those three techniques to improve accuracy performance.

b) Six phase evaluation phase: To evaluate ML, DL, and our proposed hybrid classification approaches algorithms, this research applied 10-fold cross validation. This paper assessed performance of those models utilizing various evaluation measures: Accuracy (ACC) [41], F-measure [41], Precision (PRE) [41], Recall (REC) [41], Area Under the Curve (AUC) [42], and Ensemble classifier average [28].

$$Accuracy = \frac{\sum TP+\sum TN}{\sum TP+FP+TN+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F-Measure = \frac{2*Precision*Recall}{(Precision*Recall)} \quad (5)$$

$$F\ abs = \left(\frac{AUC\ non-iv}{AUC\ iv}\right) * \left(\frac{Dose\ iv}{Dose\ non-iv}\right) \quad (6)$$

$$Ensemble(AVG) = \sum \frac{1}{n}(a1 + a2 + \cdots . + an) \quad (7)$$

IV. Results and Discussion

For empirical study, ORANGE Data Mining tool utilizes a component-based, inclusive model for DM and ML users and developers. Also, this research utilizes it for ML, and DL Models purposes. It is a combination of Python-based, and NLTK library modules which perform a set of functions such as data input, pre-processing, splitting, visualization, classification, prediction, and evaluation. Classifier methods used to classify MASR dataset utilizing: ML approach which perform KNN, SVM, NB, & LR, DL approach which perform MLP-NN for ASA. In addition, this paper suggests a novel hybrid classification technique which combined from two top ML classifiers in addition to DL classifier: LR + RF +MLP-NN to enhance accuracy for classification and prediction. k-fold cross-validation was utilized with k = 10. Accuracy, F1, Precision, Recall, AUC were utilized for evaluate MASR sentiment polarity datasets.

The results are discussed separately for each evaluation criterion. Moreover, to ensure the performance of the classifiers, this paper combined various domains to test the accuracy of various ML, DL, and our proposed hybrid approach using Arabic dialect features.

1) Accuracy (AUC): Fig. 3 represents the performance of three various classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN).

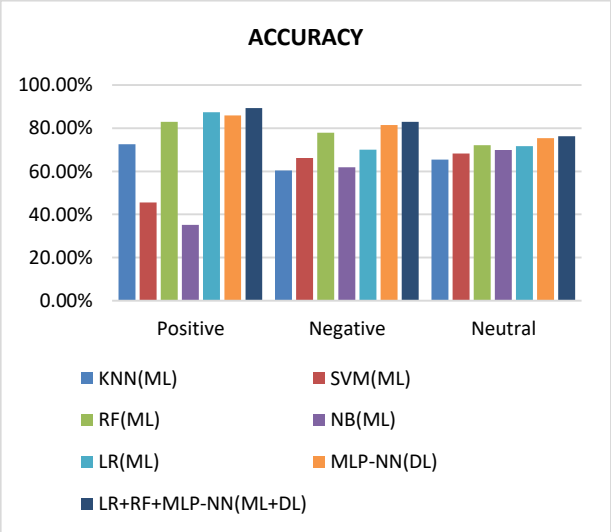


Fig. 3. Accuracy of Ml, Dl, and Hybrid (ML+DL) Approach.

[16]https://gist.github.com/mwitiderrick/363a71bc0d686383a33132aa9f896fce

[17] https://scikit-learn.org/stable/modules/feature_extraction.html

After applying ML classifiers on Positive Sentiments, results show that LR (87.5%), and RF (83%) shows better accuracy compared to a KNN (72.6%), SVM (45.5%), and NB (35.1%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that NLP-NN accuracy (86%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better accuracy (89.4%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Negative Sentiments, results mention that RF (78%), and LR (70%) shows better accuracy compared to a SVM (66.2%), KNN (60.4%), and NB (61.9%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that NLP-NN accuracy (81.5%) perform better accuracy than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better accuracy (83%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Neutral Sentiments, results mention that RF (72.1%), and LR (71.7%) shows better accuracy compared to a NB (69.9%), SVM (68.3%), and KNN (65.5%), and respectively. In addition, after applying DL classifier: MLP-NN, results observe that NLP-NN accuracy (75.4%) perform better accuracy than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better accuracy (76.3%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

2) Precision (PRE): Fig. 4 represents the various precision results of three different classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN).

After applying ML classifiers on Positive Sentiments, results mention that LR (91%) and RF (66.3%) shows better Precision results compared to a KNN (48.8%), SVM (31.8%), and NB (29%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Precision of LR (91%) perform better than Precision of NLP-NN (70.8%). And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that Precision result is (84.4%). So, LR performs better Precision result than our proposed hybrid approach.

After applying ML classifiers on Negative Sentiments, results mention that NB (96.3%) and SVM (73.9%) shows better Precision results compared to a RF (71.8%), LR (59.6%), and KNN (54.1%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Precision of NB (96.3%) perform better than Precision of MLP-NN (76.1%). And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that Precision result is (79.1%). So, NB performs better Precision results than our proposed hybrid approach, and MLP-NN.

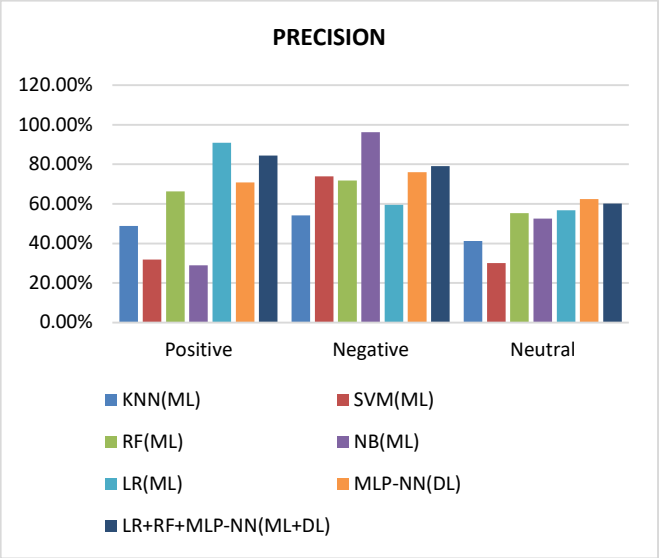


Fig. 4. Precision of Ml, Dl, and Hybrid (ML+DL) Approach.

After applying ML classifiers on Neutral Sentiments, results mention that LR (56.8%) and RF (55.3%) shows better Precision results compared to a NB (52.6%), KNN (41.3%), and SVM (30.1%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Precision of NLP-NN (62.5%) perform better than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that Precision result is (60.2%). So, MLP-NN(DL) performs better Precision results than our proposed hybrid approach.

3) Recall (REC): Fig. 5 illustrates the various recall results of three different classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN).

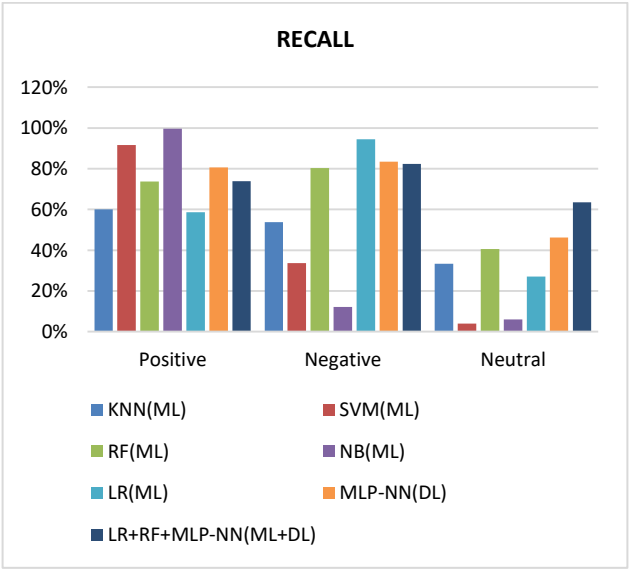


Fig. 5. Recall of Ml, Dl, and Hybrid (ML+DL) Approach.

After applying ML classifiers on Positive Sentiments, results mention that NB (99.7%), and SVM (91.6%) shows better Recall results compared to a RF (73.7%), KNN (60%), and LR (58.7%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Recall of MLP-NN (80.6%). And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that Recall of is (73.8%). So, NB and SVM perform better recall results than DL (MLP-NN) and our proposed hybrid approach (LR+RF+MLP-NN).

After applying ML classifiers on Negative Sentiments, results mention that LR (94.5%) and RF (80.3%) shows better Recall results compared to a KNN (53.8%), SVM (33.6%), and NB (12.2%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Recall of MLP-NN (83.5%). And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that Recall of is (82.3%). So, LR performs better recall results than DL (MLP-NN), our proposed hybrid approach (LR+RF+MLP-NN).

After applying ML classifiers on Neutral Sentiments, results mention that) RF (40%) and KNN (33.3%) shows better Recall results compared to a LR (27.1%), NB (6%), and SVM (4%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Recall of NLP-NN (46.2%) perform better recall than top two ML classifiers RF, KNN. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better accuracy (63.5%) than the top three classifiers: ML (RF, KNN), and DL (MLP-NN).

4) F1-Measure: Fig. 6 represents the performance of three different classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN).

After applying ML classifiers on Positive Sentiments, results mention that LR (71.3%) and RF (69.8%) shows better F1-Measure results compared to a KNN (53.8%), SVM (47.2%), and NB (45%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that F1-Measure of NLP-NN (75.4%) perform better than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better F1-Measure results (78.8%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Negative Sentiments, results mention that RF (75%), and LR (73.1%) shows better F1-Measure results compared to a KNN (54%), SVM (46.2%), and NB (21.6%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that F1-Measure of NLP-NN (79.6%) perform better than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better F1-Measure results (80.1%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Neutral Sentiments, results mention that RF (46.8%), KNN (36.9%) and LR (36.7%) shows better F1-Measure results compared to a SVM (6.7%), and NB (10%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that F1-Measure of NLP-NN (53.1%) perform better than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better F1-Measure results (61.8%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

5) Area Under the Curve (AUC): Fig. 7 represents a graph of the various AUC results of three different classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN).

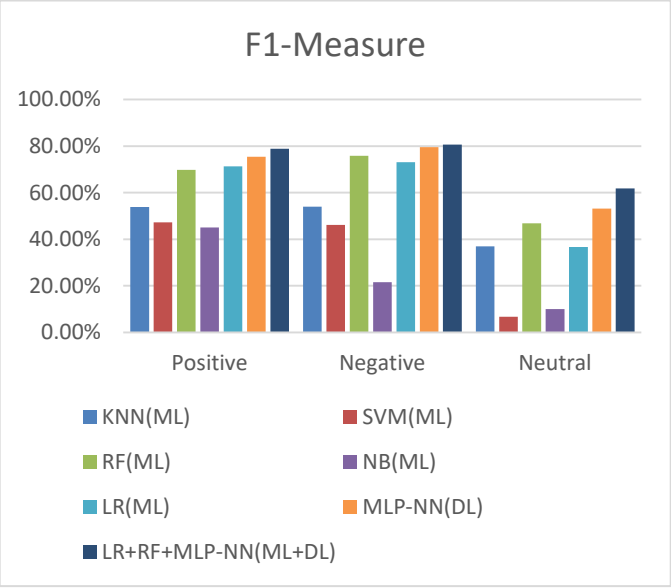


Fig. 6. F1-Measure of Ml, Dl, and Hybrid (ML+DL) Approach.

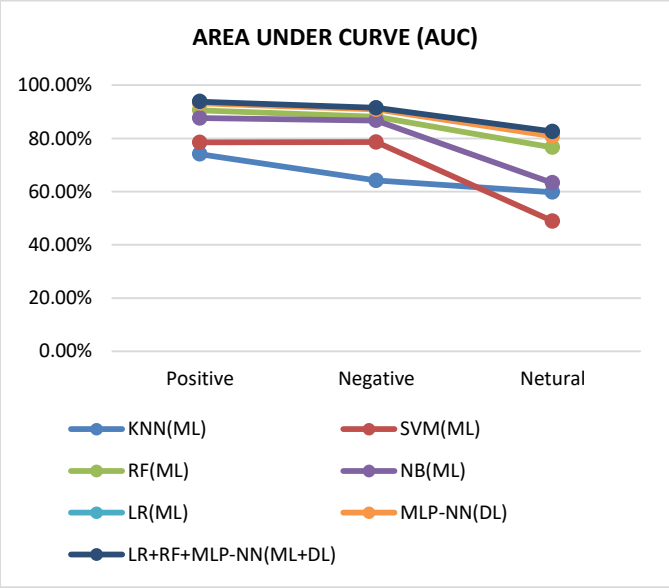


Fig. 7. AUC of Ml, Dl, and Hybrid (ML+DL) Approach.

After applying ML classifiers on Positive Sentiments, results mention that LR (93.5%) and RF (90.6%) shows better AUC results compared to a NB (87.6%), SVM (78.5%), and KNN (74.1%), and respectively. In addition, after applying DL classifier: MLP-NN, results observe that AUC of MLP-NN (93.22%) is approximate to ML classifier LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better AUC (93.8%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Negative Sentiments, results mention that LR (91.1%) and RF (88.1%) shows better AUC results compared to a NB (86.7%), SVM (78.6%), and KNN (64.2%), and respectively. In addition, after applying DL classifier: MLP-NN, results observe that AUC of NLP-NN (90.9%) is approximate to ML classifier LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better AUC (91.5%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Neutral Sentiments, results mention that LR (81.7%) and RF (76.6%) shows better AUC results compared to a NB (63.3%), KNN (59.8%), and SVM (48.9%), and respectively. In addition, after applying DL classifier: MLP-NN, results observe that AUC of NLP-NN (80.8%) is approximate to ML classifier LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better AUC (82.6%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

6) Ensemble classifier averaging: Fig. 8 represents the average performance of three various classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN) utilizing various evaluation criteria.

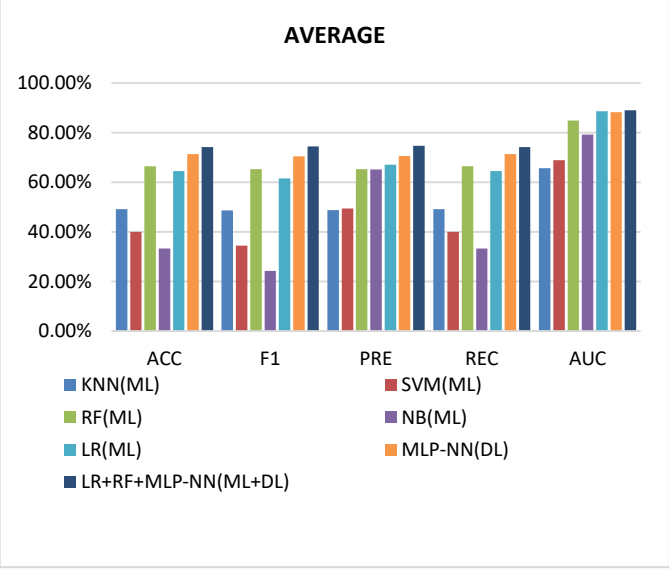


Fig. 8. Average of Ml, Dl, Hybrid (ML+DL) Approach and Evaluation Metrics.

Accuracy. After applying ML classifiers, results mention that LR (72%), and RF (70%) shows better accuracy compared to a SVM (50%), KNN (49%), and NB (45%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that NLP-NN accuracy (69%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better accuracy (74.2%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

Precision. After applying ML classifiers, results mention that LR (71.8%), and RF (69.6%) shows better precision results compared to a NB (62.6%), SVM (53.9%), and KNN (51.4%) respectively. In addition, after applying DL classifier: MLP-NN, results observe that precision of NLP-NN (68.2%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better precision results (74.9%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

Recall. After applying ML classifiers, results mention that LR (72.3%), and RF (70.3%) shows better recall results compared to a SVM (50.1%), KNN (49%), and NB (45.2%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that the recall of NLP-NN (69.1%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better recall results (74.2%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

F1-Measure. After applying ML classifiers, results mention that LR (71.8%), and RF (69.6%) shows better F1-Measure results compared to a KNN (47.9%), SVM (45.7%), and NB (41.8%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that F1-Measure of NLP-NN (68.2%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better F1-Measure results (74.2%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

AUC. After applying ML classifiers, results mention that LR (88.9%), and RF (872%) shows better result of AUC compared to a NB (829%), SVM (72.8%), and KNN (70%) respectively. In addition, after applying DL classifier: MLP-NN, results observe that AUC of NLP-NN (86.1%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better AUC (89.6%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

Finally, researchers summarize our performance for a proposed hybrid (LR+RF+MLP-NN) approach results as follow:

- Positive polarity: performs higher performance results in the following evaluation criteria: ACC (89.4%), F1 (78.8%), and AUC (93.8%).

- Negative polarity: performs higher performance results in the following evaluation criteria: ACC (83%), F1 (80.7%), and AUC (91.5%).
- Neutral polarity: performs higher performance results in the following evaluation criteria: ACC (76.3%), REC (63.5%), F1 (61.8%), and AUC (82.6%).
- Average: performs higher performance results in the following evaluation criteria: ACC (74.3%), PRE (74.8%), REC (74.3%), F1 (74.5%), AUC (89.1%).

TABLE IV. COMPARISON BETWEEN STATE-OF-ARTS HYBRID MODELS AND OUR HYBRID MODEL

Study	Dataset	Hybrid Models	Accuracy
Heikal et al. [19]	ASTD	CNN + LSTM	65.05%
Al-Twairesh et al. [20]	SemEval	SF+ GE + ASEH	80.36%
Al-Azani et al. [43]	ASTD	SGD + SGD + NuSVC	85.28%
Basir et al. [44]	COVID	CNN+ BiGRU + FastText	85.4%
Saleh et al. [38]	AJGT	LR+CBOW	86.11%
Mohammed et al. [21]	Arabic tweets	LSTM+Augmented	88.05%
Our Hybrid Approach	MASR	LR+RF+MLP-NN	89.4%

In Table IV, a comparison between the performance of our model accuracy and state-of-arts hybrid models on the various Arabic datasets (SemEval, ASTD, COVID datasets, AJGT) is presented. Researchers observe the excellence of our proposed hybrid model approach compared to the previous works.

V. CONCLUSION

This paper aims to collect a simple dataset of Mobile Apps Arabic Slang Reviews (MASR) which focus on Egyptian Arabic Slang for sentiment analysis purposes. In addition, propose a hybrid supervised classification approach which combine ML, and DL approaches to automatically predict user requirements evolution to help developers update new versions. In ML approach, apply a LR which considered a statistical method, and RF which considered a decision tree method. In DL approach, apply MLP-NN which considered a linear and non-linear method. This paper utilized various evaluation metrics like: accuracy, f-measure, recall, precision, AUC, and ensemble classifier averaging. Results show that our proposed hybrid supervised classification approach achieves good performance results in the following:

- In Positive polarity, ACC (89.4%), F1 (78.8%), and AUC (93.8%).
- In Negative polarity, ACC (83%), F1 (80.7%), and AUC (91.5%).
- In Neutral polarity, ACC (76.3%), REC (63.5%), F1 (61.8%), and AUC (82.6%).
- In Average, ACC (74.3%), PRE (74.8%), REC (74.3%), F1 (74.5%), AUC (89.1%).

A limitation in this research is the size of the dataset because it focuses only on Egyptian Arabic Slang mobile reviews. However, it considered a contribution because till now no studies concentrate on it.

VI. FUTURE WORK

In future, researchers intend to accomplish various researches in various points:

1) Apply our proposed hybrid supervised approach for automatically classify Mobile Apps categories.

2) Apply our proposed hybrid supervised approach for different Mobile Apps Arabic Slang datasets in different languages.

3) Add different feature extraction methods like word embedding, and word enrichment and n-grams, also apply different tokenization, and stemming methods.

4) Propose different hybrid ML, and DL modelling approaches and compare them with our proposed approach on different Arabic Slang datasets.

5) Apply also lexicon approach in addition to MASR dataset.

6) Extract functional, and Non-Functional, and Sentimental requirements from MASR datasets using Topic Modeling approach.

REFERENCES

[1] Fuad and M. Al-Yahya, "Analysis and Classification of Mobile Apps Using Topic Modeling: A Case Study on Google Play Arabic Apps.," Complexity, vol. 2021, 2021.

[2] I. Malavolta, S. Ruberto, T. Soru and V. Terragni, "Hybrid mobile apps in the google play store: an exploratory investigation," in Proceedings of the 2nd ACM International Conference on Mobile Software Engineering and Systems, Florence, Italy, 2015.

[3] G. Berardi, A. Esuli, T. Fagni and F. Sebastiani, "Multi-store metadata-based supervised mobile app classification," in In Proceedings of the 30th Annual ACM Symposium on Applied Computing, 2015.

[4] K. Elshakankery and M. F. Ahmed, "HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis," Egyptian Informatics Journal, vol. 20, no. 3, pp. 163-171, 2019.

[5] Y. Tian, M. Nagappan, D. Lo and A. E. Hassan., "What are the characteristics of high-rated apps? a case study on free android applications," in In Proceedings of the 2015 IEEE international conference on software maintenance and evolution (ICSME), 2015.

[6] W. Martin, App store analysis for software engineering, UK, , London: University College London, 2017.

[7] A. Finkelstein, M. Harman, Y. Jia, W. Martin, F. Sarro and Y. Zhang, "Investigating the relationship between price, rating, and popularity in the Blackberry World App Store," Information and Software Technology, vol. 87, pp. 119-139.

[8] A. Finkelstein, M. Harman, Y. Jia, . F. Sarro and Y. Zhang, "Mining App Stores: Extracting Technical, Business and Customer Rating Information for Analysis and Prediction," Research Note RN/13/21, 2013.

[9] E.-Y. Jung, C. Baek and &. J.-D. Lee, "Product survival analysis for the App Store," Marketing Letters, vol. 23, no. 4, pp. 929-941, 2012.

[10] M. Harman, Y. Jia and Y. Zhang, "App store mining and analysis: MSR for app stores," in In Proceedings of the 2012 9th IEEE working conference on mining software repositories (MSR), 2012.

[11] A. Hameed, H. A. Ahmed and N. Z. Bawany, "Survey, analysis and issues of Islamic Android apps," Elkawnie: Journal of Islamic Science and Technology, vol. 5, no. 1, pp. 1-15, 2019.

[12] M. Al-Shamani, M. Al-Sarem, F. Saeed and W. Almutairi, "Designing an Arabic Google Play Store User Review Dataset for Detecting App

Requirement Issues," in In Advances on Smart and Soft Computing, Singapore, Springer, 2022, pp. 133-143.

[13] S. R. El-Beltagy and A. Ali, "Open Issues in the Sentiment Analysis of Arabic Social Media: A Case Study," in Proceedings of the 9th International Conference on Innovations in Information Technology (IIT), 2013.

[14] B. Fu, J. Lin, L. Li, C. Faloutsos, J. Hong and N. Sadeh, "Why people hate your app: making sense of user feedback in a mobile app store," in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, 2013.

[15] M. Gómez, R. Rouvoy, M. Monperrus and L. Seinturier, "A Recommender System of Buggy App Checkers for App Store Moderators," in Gomez, Maria, et al. "A recommender system of buggy app checkers for app store moderators." 2015 2nd ACM International Conference on Mobile Software Engineering and Systems., 2015.

[16] N. Chen, S. CH Hoi, S. Li and X. Xiao, "SimApp: A framework for detecting similar mobile applications by online kernel learning," in In Proceedings of the eighth ACM international conference on web search and data mining, 2015.

[17] M. LU, and P. LIANG, "Automatic classification of non-functional requirements from augmented app user reviews," in Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering, 2017.

[18] S. Abuelenin, S. Elmougy and E. Naguib, "Twitter sentiment analysis for arabic tweets," in Proceedings of International conference on advanced intelligent systems and informatics, Cham, , 2017.

[19] M. Heikal, M. Torki and . N. El-Makky, "Sentiment analysis of Arabic tweets using deep learning," Procedia Computer Science, vol. 142, pp. 114-122, 2018.

[20] N. Al-Twairesh and H. AL-Negheimish, "Surface and Deep Features Ensemble for Sentiment Analysis of Arabic Tweets," IEEE Access, vol. 7, pp. 84122-84131, 2019.

[21] A. Mohammed and R. Kora, "Deep learning approaches for Arabic sentiment analysis," Social Network Analysis and Mining, vol. 9, no. 1, pp. 1-12, 2019.

[22] R. E. Saudy, E. S. Nasr, A. E. D. M. El-Ghazaly and M. H. Gheith, "Use of Arabic sentiment analysis for mobile applications' requirements evolution: trends and challenges," in Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, Cham, 2017.

[23] R. E. Saudy, E. S. Nasr, A. E. D. M. El-Ghazly and M. H. Gheith, "A Comparative Framework for Arabic Sentiment Analysis Research," in The 54thAnnual Conference on Statistics, Computer Sciences and Operation Research, Egypt, 2019.

[24] W. Medhat, A. Yousef and H. Korashy, "Egyptian dialect stopword list generation from social network data," The Egyptian Journal of Language Engineering, vol. 2, no. 1, pp. 43-55, 2015.

[25] M. M. Al-Tahrawi, "Arabic Text Categorization Using Logistic Regression," International Journal of Intelligent Systems and Applications, vol. 7, no. 6, p. 71, 2015.

[26] M. Al-Omari, "logistic regression optimisation for Arabic customers' reviews," International Journal of Business Intelligence and Data Mining, vol. 20, no. 3, pp. 251-273, 2022.

[27] R. Ismail, M. Omer, M. Tabir, N. Mahadi and I. Amin, "Sentiment analysis for Arabic dialect using supervised learning," in In Proceedings of the International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2018.

[28] A. Hawalah, "A Framework for Arabic Sentiment Analysis Using Machine Learning Classifiers," Journal of Theoretical and Applied Information Technology, vol. 97, no. 17, pp. 4478-4489, 2019.

[29] J. O. Atoum and M. Nouman, "Sentiment analysis of Arabic Jordanian dialect tweets," International Journal of Advanced Computer Science and Applications, vol. 10, no. 2, pp. 256-262, 2019.

[30] A. Alnawas and A. Nursal , "Sentiment analysis of Iraqi Arabic dialect on Facebook based on distributed representations of documents," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 18, no. 3, pp. 1-17, 2019.

[31] R. M. Duwairi and I. Qarqaz, "A framework for Arabic sentiment analysis using supervised classification.," International Journal of Data Mining, Modelling and Management, vol. 8, no. 4, pp. 369-381, 2016.

[32] M. Alassaf and A. M. Qamar, "Improving sentiment analysis of Arabic tweets by One-way ANOVA," Journal of King Saud University-Computer and Information Sciences, vol. 1, no. 0, pp. 1-11, 2020.

[33] A. S. AL-Jumaili, "A hybrid method of linguistic and statistical features for Arabic sentiment analysis," Baghdad Science Journal, vol. 17, no. 1, 2020.

[34] A. K. Al-Tamimi, A. Shatnawi and . E. Bani-Issa, "Arabic sentiment analysis of YouTube comments," in In Proceedings of the IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Aqaba- Jordan, 2017.

[35] M. E. M. Abo, N. Idris, R. Mahmud, A. Qazi, A. T. H. Ibrahim , J. Zubairu Maitama, U. Naseem, S. K. Khan and S. Yang, "A Multi-Criteria Approach for Arabic Dialect Sentiment Analysis for Online Reviews: Exploiting Optimal Machine Learning Algorithm Selection," Sustainability, vol. 13, no. 18, pp. 1-20, 2021.

[36] S. Bessou and R. Aberkane, "Subjective Sentiment Analysis for Arabic Newswire Comments," Journal of Digital Information Management (JDIM), vol. 17, no. 5, pp. 289-295, 2019.

[37] A. A. Sayed., E. Elgeldawi, Z. M. Alaa and G. R. Ahmed, "Sentiment Analysis for Arabic Reviews using Machine Learning Classification Algorithms," in In Proceedings of the International Conference on Innovative Trends in Communication and Computer Engineering (ITCE, 2020.

[38] H. Saleh, S. Mostafa , . A. Alharbi, . S. El-Sappagh and T. Alkhalifah, "Heterogeneous Ensemble Deep Learning Model for Enhanced Arabic Sentiment Analysis," Sensors, vol. 22, pp. 1-28, 2022.

[39] S. Alhumoud, "Arabic sentiment analysis using deep learning for covid-19 twitter data," International Journal of Computer Science and Network Security, vol. 20, no. 9, pp. 132-138, 2020.

[40] A. Elhawil, Y. Trabelsi and M. Mahfoud, "Comparison between the NB and SVM methods for multiclass Arabic sentiment analysis," in In Proceedings of the IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA, 2021.

[41] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, , informedness, markedness and correlation," arXiv preprint arXiv:2010, vol. 16061, pp. 37-63, 2020.

[42] B. Yamout, Z. Issa, A. Herlopian, M. El Bejjani, A. Khalifa, A. S. Ghadieh and R. H. Habib, "Predictors of quality of life among multiple sclerosis patients: a comprehensive analysis," European Journal of Neurology, vol. 20, no. 5, pp. 756-764, 2013.

[43] S. Al-Azani and E.-S. M. El-Alfy, "Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Analysis in Short Arabic Text," Procedia Computer Science 109C, p. 359–366, 2017.

[44] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi and U. R. Acharrya, "A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets," Knowledge-Based Systems, vol. 228, 2021.

User Evaluation of UbiQuitous Access Learning (UQAL) Portal: Measuring User Experience

Nazlena Mohamad Ali[1]
Institute of IR4.0 (IIR4.0)
Universiti Kebangsaan Malaysia
Selangor, Malaysia

Wan Fatimah Wan Ahmad[2]
Computer & Information Sciences Department, Universiti Teknologi PETRONAS, Perak, Malaysia

Zainab Abu Bakar[3]
Faculty of Computer and Information Technology
Al-Madinah International University
Kuala Lumpur, Malaysia

***Abstract*—The goal of user experience (UX) research in human-computer interaction is to understand how humans interact with technology. This paper aimed to evaluate the interface and user experience of UbiQuitous Access Learning Portal (UQAL) and make recommendations for the system interface. UQAL Portal is an e-learning web portal that teaches a targeted group of users how to start a business or an online business using an e-learning portal. The portal will be used to search for business-related information, among other things. The User Experience Questionnaire (UEQ) is used to evaluate user experience. The interface is evaluated using a heuristic evaluation technique based on Nielsen's ten heuristics. According to the UEQ results, the average score for each aspect in 30 UQAL users is: Attractiveness aspect: 1.77; Perspicuity aspect: 2.20; Efficiency aspect: 2.30; Dependability aspect: 1.73; Stimulation aspect: 0.63; and Novelty aspect: 1.27. A comparison of the average score in the dataset product of UEQ Data Analysis Tool revealed that the Perspicuity, Efficiency, and Dependability aspects of UQAL belonged to the Excellent category. The Attractiveness and Novelty aspects could be categorized as Good, and its stimulation could be categorized as Below Average. Four evaluators participate in the heuristic evaluation, which tests all user categories in UQAL. The findings of this study can be used as a suggestion and reference for UQAL Portal improvement.**

***Keywords*—User experience questionnaire; user experience; user interface; heuristic evaluation**

I. Introduction

Because of the rapid evolution of digital technologies, new forms of human interaction and experiences are becoming possible. To achieve a positive user experience with technology, service providers must ensure a high user experience quality. Nowadays, users' demand for products is no longer limited to functional satisfaction but also includes psychological needs [1], which involve emotional, intellectual, and sensual aspects [2]. To date, user experience (UX) research has attempted to comprehend how humans interact with technologies such as computers, mobile phones, telecommunications networks, and other digital systems [3]. Similarly, user experience (UX) is a critical factor in the commercial success of digital products. It appears that the new UX movement is gaining traction among academics and industry practitioners who are looking for innovative approaches to improve the experiential qualities of technology use.

As a result, this paper aims to understand user experience better when interacting with technologies by measuring user experience while interacting with the UQAL Portal. UQAL is an abbreviation for UbiQuitous Access Learning. The UQAL Portal will bring a Digital Transformation for learners to access business-related information from the e-learning portal and for educators to supply business-related information into the e-learning portal. The B40 group in Malaysia is the target audience for the UQAL Portal. The B40 group represents the bottom 40% of income earners. The goal is to assist the B40 group in learning how to start a business or online business using the UQAL Portal.

Furthermore, the portal will be used as a platform for the B40 group to search for business-related information, among other things. UQAL is evaluated based on its user interface and user experience, and the interface is evaluated using a heuristic evaluation technique. A User Experience Questionnaire (UEQ) assesses UQAL's user experience. The evaluation of the user experience can provide feedback about the product or service and facilitate product improvements and acceptance among the targeted users.

The rest of the paper is structured as follows: Section II identifies the Experience Evaluation Methods (UXEMs) used to evaluate and measure user experience in previous papers. In Section III, the paper discusses UX evaluation methods on the UQAL Portal. Section IV discusses the findings, followed by the conclusion, which concludes and provides insight for the improvement and future direction of the UQAL Portal.

II. Background Work

A. User Experience (UX)

The International Organization for Standardization (ISO) 9241-110:2010 defines user experience as a person's perceptions and responses resulting from the use and anticipated use of products, systems, or services. Several studies have been conducted to explain the meaning and concept of user experiences with technology. User experience is used to stimulate the HCI (Human-Computer Interaction) research by focusing on the aspect of usability that goes beyond usability and its task-oriented instrumental values [4]. According to Vermeeren et al. [5], user experience examined how an individual felt about using a product, i.e., the experiential, affective, essential, and beneficial aspects. According to Melançon et al. [6], when interacting with a

product or service, the user experience was described as a fleeting, primarily evaluative feeling (good-bad), and it was about having a positive experience through a system. Lipp [7] emphasizes that user experience is subjective because it is about an individual's performance, satisfaction, feelings, and thoughts about a product or service. Despite the lack of a clear definition, the concept of user experience has emerged as an important design consideration for interactive systems [8]. According to Allam, Razak and Hussin [9], user experience is dynamic and involves multiple research areas, including HCI, product design and development, and psychology. As a result, user experience can be viewed as a phenomenon, field of study, or practice. Some work on measuring user experience and usability was carried out by [10] [11] [12] [13]. These studies assess user interaction and product usage, including satisfaction.

The user experience is dynamic because it changes over time as conditions change. As a result, user experience should be valuable after interacting with an object and before and during the interaction. While evaluating short-term experiences is important, given the dynamic changes in user goals and needs resulting from contextual factors, it is also critical to understand how (and why) experiences evolve [5]. A product's effect on a user is called the user experience. In addition, Türkyilmaz, Kantar, Bulak and Uysal [14] stated that user experience is an emotional interaction that begins with usage as a feeling. It is about how we feel and remember after using the product. The term "user experience" refers to using a device to create an experience rather than just creating a fancy interface.

Although there is no agreement in the literature on defining user experience, everyone agrees that it is a complex concept and should not be confused with usability or user interface [15]. Hellweger and Wang [15] conducted a thorough examination of the user experience concept and proposed a user experience conceptual framework. There are numerous perspectives on user experience, and it is understood in various ways by various disciplines and can be viewed from various perspectives [16]. User experience can be academically defined as any aspect of a user's interaction with a product, service, or company [17]. Nonetheless, user experience is regarded as desirable. However, what something exactly means is still up for debate, and it is a highly interdisciplinary topic [18].

A large and growing body of literature has been devoted to understanding user experience (UX) better. Due to the variety of concepts and the flexibility of adding and removing them when stating a definition, it is not easy to have a unique and general definition for user experience. User experience, in our opinion, is primarily associated with the overall design and presentation of online software solutions such as websites or apps. To date, the analysis appears to have focused on user experience in specific domains and fields. For instance, user experience evaluations in games and interactive entertainment [8], [19], [20], [21], culture [22], [23], [24], robotic [25], safety-critical domains [26], and in business and management [18] and [27].

User experience evaluations in games, and more broadly in interactive entertainment systems, had previously been performed over the last ten years [19]. HCI user experience evaluation methods are used during game development to improve user experience. To better understand the concept of user experience, HCI borrowed and explored aspects of the gaming experience such as immersion, fun, and flow [19]. Nagalingam and Ibrahim [21] conducted additional research on the user experience elements for the evaluation and design of educational games (EG). It is critical to identify the appropriate elements to model the right user experience framework for EG to assist the designer in producing an effective educational game [21].

Several studies have been conducted to investigate user experience with social robots. In 2017, Alenljung, Andreasson, Billing, Lindblom and Lowe [25] demonstrated how the user interacted with the humanoid robot Nao while conveying emotions to the robot through touch. The research objective was to gain a better scientific understanding of affective tactile interaction and see if theories and findings from emotional touch in user experience could be applied for future robotic technologies [25]. It was preliminary to conduct additional user experience studies in the Human-Robot Interaction research area.

Grundgeiger, Hurtienne, and Happel [26] recently emphasized the importance of the personal experience of consumers in security-critical domains who engage with technology such as healthcare. They summarized "interaction" concepts based on modern theories of HCI, which include personal user experience as an essential construct. They concluded that improving user experience could improve technology design, employee well-being, and modern safety management [26].

Luther, Tiberius and Brem [18] recently conducted a bibliometric analysis to identify the evolution of scientific research on user experience between 1983 and 2019. However, despite its importance for competitiveness, customer satisfaction, customer retention, and, ultimately, firm performance, the topic has so far been discussed in the HCI field rather than in business and management. As a result, businesses must adopt a successful user experience approach [18]. It is consistent with Erdos's [27] research, which found that user experience is one of the most important determining factors in the case of business software products and services. They recommended that future research concentrate on business and management-related topics.

III. Materials and Methods

The evaluation methods for user experience are another path for undergoing user experience studies. The primary goal of evaluating user experience is to support and aid in selecting the best design, ensure that development is on track, or measure and clarify whether the final product meets and exceeds the initial user experience targets [9].

There are an increasing number of methods for assessing user experience available at all stages of the development process. Several studies attempted to conduct a comprehensive review of user experience evaluation methods to understand the available methods better. Surveys on these contributions are already available [5], [28], and [29]. A study by Vermeeren et al [5] had discovered 96 user experience evaluation methods

both from academia and industry. They also discovered a need for development of UX evaluation methods, such as early-stage methods, methods for social and collaborative UX evaluation, and establishing practicability and scientific quality.

Bargas-Avila and Hornbk [28] conducted an integrated review of user experience, looking for similarities across products, experience dimensions, and methodologies (time frame restricted to 2005–2009). According to the study's findings, questionnaires (self-developed questionnaires) were the most commonly used method of assessing user experience. In addition, qualitative methods included semi-structured interviews, focus groups, open interviews, user observation, video recording analysis, and diary analysis. However, psychophysiology is rarely used to improve user experience [28]. Table I summarizes the data collection methods used by Bargas-Avila and Hornbk [28].

Maia and Furtado [29] conducted a systematic review on user experience evaluation (time frame restricted to 2010–2015). According to Maia and Furtado [29], most of the studies used questionnaires to assess the user experience rather than other tools and techniques such as interview, observation, reports, video recording, eye-tracking, etc. They reported that psychophysiological analysis was not yet used in user experience evaluation models because most studies evaluated the user experience manually. According to literature reviews, many different types of user experience evaluation methods are available in the industry and academia. However, methodological improvements in evaluating user experiences that focus on product use and their specific needs such as development phase, type of experience addressed, target users, and evaluation objective are required.

A. Respondents

Respondents were found through a WhatsApp Group announcement. Users who wish to participate in this survey have received an invitation to do so. All respondents had been informed about the survey's objectives and methods. The invitation contained a link to our survey, which was created using the online survey tool Google Forms.

B. Data Analysis

The data gathered during the evaluation process is both quantitative and qualitative. The open-ended questionnaire yields qualitative data. The UEQ provided the quantitative data. The results of the evaluation are then summarized into a table. The data was then analyzed to determine the user experience level of UQAL. The system's user experience is graded on six scales: Stimulation, Perspicuity, Efficiency, Dependability, Attraction, and Novelty. The level of user experience for each scale is calculated by processing statistical data with the UEQ Analysis Data Tool. After obtaining the score for each scale, the data is displayed using a benchmark graph to determine the quality of UQAL in comparison to other products in the data set UEQ Analysis Data Tool.

C. UQAL Portal Interface

Evaluation is a stage where the UQAL Portal's effectiveness and efficiency are perceived. The user's interface effect is measured, which concerns how simple the portal can be learned, its usability and user experience, and problems that may occur on the portal are identified. UQAL is evaluated based on its user interface and user experience. This evaluation aims to measure the user experience and user interface when interacting with the portal. A heuristic evaluation technique is used to evaluate UQAL's user interface. According to Nielsen [30], a heuristic evaluation is carried out by a group of evaluators who are given an interface. They are then asked to evaluate whether each element adheres to a set of established heuristic uses.

UQAL Portal is an e-learning web portal that teaches a targeted group of users how to start a business or an online business using an e-learning portal. The UQAL Portal can be found at https://yutp-uqal.com/. The B40 group in Malaysia is the target audience for the UQAL Portal, and the B40 group represents the bottom 40% of income earners. UQAL Portal will bring a Digital Transformation for learners to obtain business-related information from the e-learning portal and for educators to provide business-related information to the e-learning portal. The user will interact with the e-Learning portal through GUI elements such as menus, buttons, checkboxes, search fields, pagination, and notification. Fig. 1–3 depicts the UQAL interface's main menu.

TABLE I. EXAMPLES OF DATA COLLECTION METHODS

Data Collection methods	Examples
Questionnaires	SAM scale: user feedback assessed with a self-developed questionnaire; AttrakDiff; Lavie & Traktinsky; Other surveys (e.g., FSS, IMI, Emocards).
Interviews (semi-structured and open)	Interview regarding interaction experience; engagement; to understand the enchantment.
User observation (live)	In-situ observation of apps usage; observation of people experience using apps.
Video recordings	Recordings of interactions with apps; videos to capture listening experiences on the apps.
Focus groups	Group discussion to investigate preferences.
Diaries	Emotions assessed with diaries; diaries using day reconstruction & experience narration.
Probes	Participants were given a probe kit with a brief personal explanation and instruction.
Body movements	The choreography of interaction with apps was evaluated by analyzing the movements.
Psychophysiological measures	Psychophysiology (galvanic skin response, EMG, heart rate).

Fig. 1. Main page of UQAL Portal.

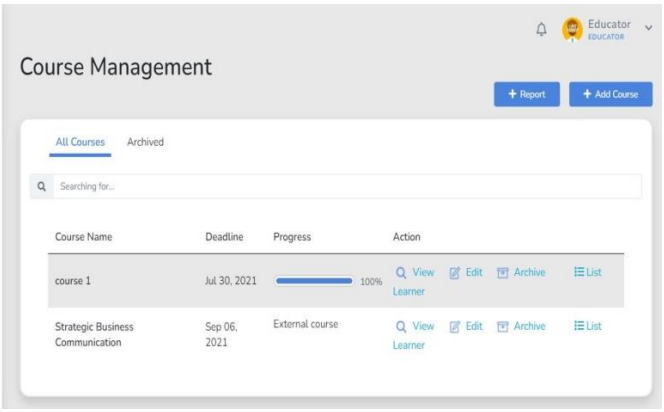


Fig. 2. Course Management Menu for Educator Interface.

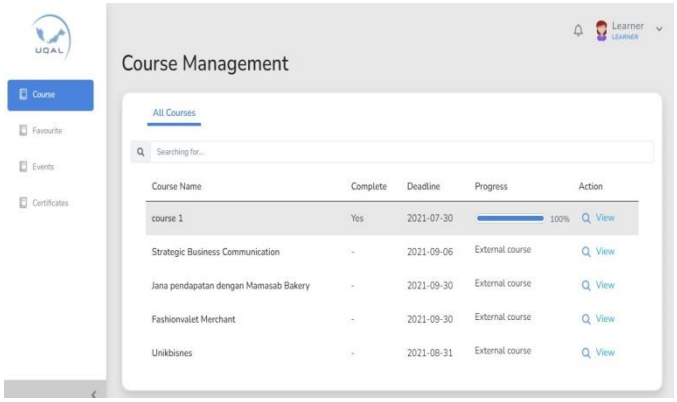


Fig. 3. Course Management Menu for Learner Interface.

D. Instruments

1) User Experience Questionnaire (UEQ): The method was chosen for this study. The questionnaire was divided into four sections. The first section asked a few questions about the user's demographic information (i.e., age, gender, race, occupation, working experiences). Users rate the usability evaluation, including the portal interface, ease of use, and learnability. These sections used a five-point Likert scale with 1 (Strongly Disagree), 2 (Disagree), 3 (Neutral), 4 (Agree), and 5 (Strongly Agree) was employed. The UEQ in the third section is used to assess the user experience of the UQAL e-learning portal. The UEQ can be accessed for free and is available at https://www.ueq-online.org/. The UEQ has seven scales and 13 items in total (as shown in Fig. 4). This study employed only the 13 items of UEQ related to the user experience to cover the user's psychological aspects such as feelings of pleasure, disappointment, and stimulation when using the portal interface. Table II shows each of these scales in detail. This section allows users to choose their own experiences and opinions while interacting with the portal. Finally, we ask the user to provide any comments or suggestions for the portal's improvement for the open-ended questions.

	1	2	3	4	5	6	7	
Boring	o	o	o	o	o	o	o	Exciting
Not interesting	o	o	o	o	o	o	o	Interesting
Difficult to use	o	o	o	o	o	o	o	Easy to use
Complicated	o	o	o	o	o	o	o	Easy
Inefficient	o	o	o	o	o	o	o	Efficient
Impractical	o	o	o	o	o	o	o	Practical
Does not Meet Expectations	o	o	o	o	o	o	o	Meets Expectations
Demotivating	o	o	o	o	o	o	o	Motivating
Cluttered	o	o	o	o	o	o	o	Organized
Inferior	o	o	o	o	o	o	o	Valuable
Unattractive	o	o	o	o	o	o	o	Attractive
Dull	o	o	o	o	o	o	o	Creative
Unpleasant	o	o	o	o	o	o	o	Pleasant

Fig. 4. User Experience Questionnaire (UEQ) Items.

TABLE II. USER EXPERIENCE QUESTIONNAIRE (UEQ) ASPECTS

Aspects	**Items**
1.Attractiveness General impression of the product. Do users like or dislike the product?	unattractive/attractive, unpleasant/pleasant
2. Perspicuity Is it easy to understand how to use the portal? Is it easy to get familiar with the portal?	difficult to learn/easy to learn, complicated/easy
3. Efficiency Is it possible to use the product fast and efficiently? Does the user interface look organized?	inefficient/efficient, impractical/practical, cluttered/organized,
4. Dependability Does the user feel in control of the interaction? Is the interaction with the product secure and predictable?	does not meet expectations/ meets expectations
5. Stimulation Is it interesting and exciting to use the portal? Does the user feel motivated to further use the portal?	boring/exciting, not interesting /interesting, demotivating/motivating, inferior/valuable
6. Novelty Is the design of the portal innovative and creative? Does the portal grab attention?	dull/creative

2) Heuristic Evaluation

a) Nielsen's ten heuristic principles are described below:

1) Visibility of system status: This system should always keep users informed of what is going on by providing appropriate feedback in a timely manner.

2) Match between system and the real world: The system should speak the user's language, using words, phrases, concepts that the user is familiar with, and adhere to real world conventions rather than system-oriented terms.

3) User control and freedom: Users frequently select system functions by accident, necessitating a marked "emergency exit" to exit the undesirable state without going through an extended dialogue.

4) Consistency and standards: Users should not guess whether various words, situations, or actions mean the same thing. Observe platform conventions.

5) Error prevention: A careful design that prevents a problem from occurring in the first place is even better than good error messages. Either eliminate error-prone conditions or check for them and provide users with a confirmation option before proceeding with the action.

6) Recognition rather than recall: Make objects, actions, and options visible to reduce the user's memory load. The user should not have to recall information from one section of the dialogue to the next. When appropriate, system instructions should be visible or easily accessible.

7) Flexibility and efficiency of use: Unseen accelerators may frequently speed up the interaction for the expert user, allowing the system to cater to both inexperienced and experienced users. Allow users to personalize frequently performed actions.

8) Aesthetic and minimalist design: Dialogues should not include irrelevant or used infrequently. Every additional unit of information in a conversation competes with the relevant information units, reducing their relative visibility.

9) Help users recognize, diagnose, and recover from errors: Error messages should be written in plain language (no codes), accurately describe the problem, and constructively suggest a solution.

10)Help and documentation: Even though it is preferable if the system can be used without documentation, assistance and documentation may be required. Any such information should be easy to find, focused on the user's task, list concrete steps to be taken, and not be too large.

b) Data Collection Procedures: The following are the data collection steps in the heuristic evaluation:

- Step 1: Establish an appropriate list of heuristics. This survey used the model based on Nielsen's 10 heuristics.
- Step 2: Identify 3 to 4 evaluators (experts). They were assuring their knowledge of the relevant industry. Experts were defined in this survey as people with several years of job experience in the software and information technology fields.
- Step 3: Briefing the evaluator/expert. They inform the evaluator about what they are expected to do and cover during their evaluation. The evaluator has explained the scope and objective of the portal inspection and the characteristics of the portal's users.
- Step 4: Evaluation phase. Evaluators must have free access to the portal to identify elements to analyze. Individual elements are examined by evaluators using heuristics. They also investigate how these fit into the overall design, meticulously documenting all issues encountered.
- Step 5: Report issues/problems. Evaluators complete the questionnaire given and report any issues and problems they discover. The evaluator's task at this stage is to assess the list of 10 Usability Heuristics for User Interface Design [30] in Table III.

TABLE III. NIELSEN'S 10 HEURISTICS

Heuristics	**Yes/No**	**Comment/ Remark**
1. Visibility of system status		
2. Match between systems and the real world		
3. User control and freedom		
4. Consistency		
5. Prevent Errors		
6. Recognition rather than recall		
7. Flexibility and efficiency of use		
8. Aesthetic and minimalist design		
9. Help users recognize, diagnose, and recover from errors.		
10. Help and Support		

The data obtained from this technique is a list of interface problems based on the evaluators' heuristic principles. The evaluation results are then compiled into a table that provides a detailed breakdown of the issues and recommendations.

IV. RESULTS AND DISCUSSION

A. Demography

Thirty users participated in the user experience survey (19 females, 11 males). Most respondents were between the ages of 18 and 35 (n = 23), followed by those between the ages of 36 and 55 (n = 6), with the remainder being over the age of 55 (n = 1). Malay (97%) and Chinese are the most common ethnic groups (3%). The majority had a bachelor's degree or were enrolled in a bachelor's degree program (70%). 7% had high school diplomas, 10% had college diplomas, and 13% had graduate degrees.

In terms of current employment status, 63% were full-time employees, 20% were students, 7% were self-employed, and 1% were full-time freelancers, unemployed, or retired. Respondents' current occupations were education and training (17%), computer and software (13%), administrator (7%), students (6%), and other fields (1%), in that order. The average working experience ranged from more than seven years (37%) to four to six years (30%), one to three years (13%), less than six months (7%), and none at all (13%).

Furthermore, approximate monthly household income for the respondent shows that 37% have more than 4500 Malaysian Ringgit, 20% have 2500–3500 Malaysian Ringgit, 10% range from 1500–2500 Malaysian Ringgit and less than 2500 Malaysian Ringgit. In addition, 7% ranges from 3500–4500 Malaysian Ringgit. The remaining respondents (17%), on the other hand, preferred not to respond. Table IV summarizes the detailed demographic information.

TABLE IV. DEMOGRAPHIC PROFILE OF RESPONDENT

Demographic Profile	Total N = 30 (%)
Age	
18– 35	23(76.7)
36–55	6(20)
>55	1(3.3)
Race	
Malay	29(96.7)
Chinese	1(3.3)
Education level	
Bachelor's degree	21(70)
Graduate degree (MS, Ph.D.)	6(13.4)
College Graduate	3(10)
High School	2(6.7)
Employment status	
Full-time employment	19(63.3)
Student	6(20)
Self-employed	2(6.7)
Working experience	
7 years or more	11(36.7)
4 to 6 years	9(30)
1 to 3 years	4(13.3)
No working experience	4(13.3)
Monthly household income	
>RM 4500	11(36.7)
RM3500–RM4500	2(6.7)
RM 2500–RM3500	6(20)
RM1500–RM2500	2(6.7)
Less than RM1500	2(6.7)

B. Usability Evaluation

This survey evaluates the portal's usability with a few items identifying general interface design and layout, ease of use, and learnability. Overall, participants gave positive feedback on usability aspects, as shown in Table V. 79.9% thought the portal interface was pleasant and easy to use (n = 24). In comparison, 83.3% thought the sequence of screens, organization of information presented, and graphical presentations were simple to understand (n = 25). As a result, 89.9% agreed that the portal was simple to use (n = 27), 86.6% agreed that it was easy to find needed information (n = 26), and 90% of respondents understood the menu (n = 26). Overall, most of the participants, 86.6%, were satisfied with the easiness of the portal (n=26). In terms of learnability, most respondents (96.6%) said that it was easy to learn how to use the portal; 89.9% said it helped them become more productive quickly. Another 93.3% found the information in the portal to be effective and helpful.

C. User Experience

Overall, the score indicates that the UQAL Portal gets a positive evaluation from users. Results from UEQ show that the overall score is in the positive range. The Likert scale data has been transformed into the UEQ Data Analysis Tool in an Excel sheet to calculate the scale means and compare the products in the benchmark data set. The measured scale means are determined by comparing them to existing values from a benchmark data set (https://www.ueq-online.org/). Comparing the results for the evaluated product with the data in the benchmark allows conclusions about the quality of the evaluated product compared to other products. Table VI shows the score of each user experience aspect.

TABLE V. MEAN AND STANDARD DEVIATION FOR EACH USABILITY ITEM

Usability Aspect	Mean and Standard Deviation, SD						
A. General Interface Design and Layout	**1**	**2**	**3**	**4**	**5**	**Mean**	**SD**
The interface of the portal is pleasant.	1	0	5	20	4	3.87	3.42
I like using the interface.	1	0	5	18	6	3.93	3.49
The sequence of screens was clear.	1	0	4	13	12	4.17	3.74
The organization of information presented was clear.	1	0	4	13	12	4.17	3.74
The graphical presentations (i.e., icons) are easy to interpret.	1	1	3	18	7	3.97	3.54
B. Ease of Use	**1**	**2**	**3**	**4**	**5**	**Mean**	**SD**
It was simple to use this portal.	0	0	3	16	11	4.27	3.79
It was easy to find information I needed.	0	0	4	15	11	4.23	3.76
It is easy to understand the functions of the menu items.	0	1	2	15	12	4.27	3.80
The information (i.e., online help, on-screen messages, and other documentation) provided in this portal is clear.	0	1	3	16	10	4.17	3.71
Whenever I make a mistake using the portal, I recover easily and quickly.	0	0	3	15	12	4.3	3.82
The portal gives error messages that clearly tell me how to fix problem.	0	1	9	16	4	3.77	3.31
Overall, I am satisfied with how easy it is to use this portal.	0	0	4	19	7	4.1	3.61
C. Learnability	**1**	**2**	**3**	**4**	**5**	**Mean**	**SD**
It was simple to use this portal.	0	0	1	16	13	4.4	3.91
I would imagine that most people would learn to use this portal very quickly.	0	0	2	18	10	4.27	3.78
I believe I became productive quickly using this portal.	0	0	3	21	6	4.1	3.61
The information provided in this portal is effective and helpful.	0	1	1	16	12	4.3	3.83
The online Help facility is useful.	0	1	3	15	11	4.2	3.74

TABLE VI. USER EXPERIENCE QUESTIONNAIRE (UEQ) RESULTS

Aspects	Average Score	Compared to Benchmark
1. Attractiveness	1.77	Good
2. Perspicuity	2.20	Excellent
3. Efficiency	2.30	Excellent
4. Dependability	1.73	Excellent
5. Stimulation	0.63	Below Average
6. Novelty	1.27	Good

The benchmark results from UEQ Data Analysis Tool revealed that perspicuity, efficiency, and dependability aspects belonged to the Excellent category, indicating that UQAL is included in the best 10% range of results, implying that 10% of the products in the dataset are better and 75% are worse. However, the stimulation aspect of UQAL could be classified as Below Average, which means that 50% of products on the dataset are better than UQAL while 25% are worse. The overall score is in the positive range, according to the evaluation of UEQ results. Minor issues on UQAL have not been shown to impact user experience significantly. Fig. 5 depicts the benchmark graph.

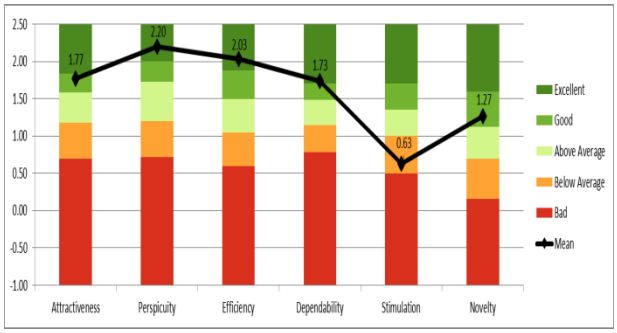


Fig. 5. UQAL's Benchmark Graph.

Another thing that this survey wanted to look into was whether UQAL should have any additional essential features. Table VII includes some comments on all the missing features mentioned.

We are looking at which aspects/features the UQAL Portal users like the least and most. According to the survey results, some respondents like how simple it is to use and understand the portal. Some respondents said they were straightforward when asked about the portal's features. For example, "I like the portal structure; it is straightforward." They commented, "This portal is simple and easy to understand." Some of the respondents commented, "User-friendly." This portal helps me find any business courses. I can easily organize and manage courses from the beginning to the end". Others commented that it is "so easy for people to understand the flow of the system because each page has different information".

Some respondents stated that the user interface design is their least favorite. They felt the portal's interface was not interesting enough to draw their attention. Some of them stated:

"The theme of the portal does not seem very interesting. Color combinations could be used to make the portal look better."

"The thing I like the least is the inconsistent type of fonts used and the size of the fonts. I found certain words or sentences do not start with a capital letter, which does not represent the professional side."

"The color of the portal. This e-learning web portal is for Malaysians who want to start a business online. The color of a website plays a vital role in attracting more B40 groups."

"The team can research which fonts are compatible for each part, especially for the Business Opportunities interface and UQAL course interface. I found certain fonts used are 'awkward', and the layout and the color of the fonts should be consistent. "

TABLE VII. USER EXPERIENCE QUESTIONNAIRE (UEQ) RESULTS

Features	Comments
1. Course/Event details	"Course introduction, course timeline, and instructors' information" "State the date when this portal updates the information of the course and event. So, the user will know the information was updated." "Can display multiple categories of data like all courses, my courses, external courses in a single screen." "Add more information of the courses." "Have a calendar to show/ list next course register." "Add description for the courses." "Have a list of courses by category."
2. Customer Service Chat/Online Chat	"Chabot or online helper to assist users when they face any issues when using the portal." "Chat feature to allow peer engagement and learners-instructor interaction." "Any online learning platform should have chat features to enable for peer engagement as well as learner-instructor interaction."
3. Information/ Content	"Give information about another interesting portal" "Can add detail grant for SME, provided from the gov-ernment" "Introduced more local corporate and business com-pany starter." "Maybe can add 'dashboard' that include information such as a graph to prove how UQAL Portal help the B40 group start the business using this portal." "Information on business events and business oppor-tunities" "Company and corporate sector involved mostly big known." "Should have "about" section which could explain to people what the portal is about."
4. User Interface Design	"Perhaps the portal should be more organized with a drop-down menu…" "No attractive colors or graphics." "Greyish button. Hope more eye-catchy." "Add more pictures or graphics to make this portal interesting." "More interesting interface maybe can add animation, the welcoming or introduction video." "It would look nicer with better images resolution."
5. Advanced Search	"Have sort and filter searching" "Allow multiple searching criteria in one screen."
6. Bilingual (BM/BI)	"Make it friendlier for example, bilingual feature for easy to understand."

D. Heuristic Evaluation Analysis

Four evaluators evaluate with backgrounds in software and information technology. Two of them have more than seven years of experience as software developers. One has over ten years of experience as an information technology administrator, and the other is a research graduate in usability. Based on the severity rating, the evaluators discussed some issues and made recommendations for improvement.

TABLE VIII. UQAL HEURISTIC EVALUATION RESULTS

Heuristic	Comments/Issues	Recommendations
1.Visibility of system status	Some courses are already marked as completed, yet in the listing, it still shows "No" under the complete column and the progress bar; I'm not sure how it works/functions (already complete, but the progress bar still shows 25%).	This portal should always keep users informed about what is going on, through appropriate feedback within reasonable time.
2. Match between systems and the real world	The portal has no elements of positive encour-agement (rewards, praise, personalization, etc.) to boost users' motivation. This type of element is essential in online learning since it requires users to learn independently.	The system should speak the user's language with words, phrases, and concepts familiar and follow real world conventions rather than system-oriented terms.
	There are no features that allow learners to interact with one another and with instructors/ lecturers/ trainers.	Any online learning platform should have chat features to enable peer engagement and learner-instructor interaction.
3. User control and freedom	When the user makes an error on a certain field, the system removes all the information that the user has filled in, even though that information is supposedly correct.	Support undo and redo.
4. Consistency	Some buttons are not suitable. For example, there is a button 'Report' I thought it was for adding a report, but it was to generate Report.	Use the standard color of buttons. Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
5. Prevent errors	The function is not working well for evaluation. When I try to add a new course as an educator, the system makes it compulsory to add the image of the banner. When I didn't add the image, it showed an error message. However, the page redirects me to the front page. So, I need to re-key the form again.	Properly test the portal to make sure all functionality is working. The system must be functioning well to be successful. Prevents a problem from occurring in the first place. Present users with a confirmation option before they commit to the action.
6. Recognition rather than recall	There is no error warning message when the user makes a mistake.	The system should prevent users from making mistakes. Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another.
7. Flexibility and The efficiency of use.	Advance search feature: Users do not get the benefit of a search menu there.	The search feature should add sorting and a filtering function.
	Its multi-platform, but it can't be used for the mobile version well.	Create a mobile-friendly web portal for users to access because not everyone has a laptop or tablet to access the portal.
	It would be better if you could add the calendar management for learners and educators to view the courses and events they join/conduct. For example, if I'm a learner and I click to join the event/course, then the event will be added to my calendar.	Make the calendar to be viewed monthly/weekly. So, that it easier to check which event/course that I have joined or to join.
	I'm not sure how the courses will be conducted. So that leaner can always come to this website to review back the provided material. When the course is already marked as complete, there is nowhere for me to view back what the courses are all about.	It would be better if an educator can up-load the teaching material (e.g., Power-Point slides or others material).
8. Aesthetic and minimalist design	The interface for "Course Management" (learner view) is not convenient to use. All the courses are displayed in one listing.	It would be better if you could display the listing for the courses that have already been completed in one tab and the courses not yet completed in another tab. Or you could just add the filter there to allow users to filter the listing.
	The portal theme appears to be uninteresting. This portal does not appear to employ vivid colors.	Color combinations could be used to improve the portal's aesthetic value.
	The image size used does not fit and not match the box provided.	It is possible to match all the pictures at the same size and clear.
	The dashboard for learners should be appealing and dynamic.	Choose dashboard UI elements carefully; otherwise, learners will become discouraged.
	The sidebar's use of repetitive icons appears to be confusing.	The button positioning should be con-sistent and clear (e.g., the Join button).
9. Help users recognize, diagnose, and recover from errors.	When the user makes an error on a certain field, the system totally removes all the information that the user has filled in, even though that information is supposedly correct.	Recovery from Error. Help user to recover if making an error. Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
10. Help and support	Lack of Help to guide users.	Put Guideline to help the user. Create a help menu to make it easier for users to use the portal.
	There is no sitemap of the portal.	Create a site map to make it easier for users to navigate the portal.

The majority of problems can be found in Aesthetic and Minimalist Design principles. Two of the evaluators noticed some issues with UQAL's interface's aesthetic. The principle issues include a colorless interface, size, images resolution, icons, and buttons that should have aesthetic values according to the evaluator. The Search and Course function menus should be improved based on flexibility and efficiency of use. Some errors occur while performing certain tasks. Table VIII displays the outcome of the heuristic evaluation.

Although heuristic evaluation revealed some significant flaws in the UQAL's user interface design, it had no direct impact on the UEQ's overall score, which was positive. Previous research on heuristic evaluation has shown that it identifies more minor usability issues in an interface than other methods [31]. Regardless, the UEQ results show that the overall score is positive. Minor issues discovered during the heuristic evaluation do not appear to significantly impact user experience on UQAL.

V. Conclusion

Finally, based on the heuristic evaluation results, the evaluators discovered some issues with Nielsen's heuristic principles on UQAL's user interface design, which are commonly found in the Aesthetic and Minimalist Design principles, as well as flexibility and efficiency of use. The outcome of heuristic evaluation is a recommendation of issues and problems that must be addressed. Nonetheless, according to UEQ results, the user experience of the UQAL Portal is adequate. The sufficient average score of each aspect demonstrates this. The results of this experiment can be used as a reference for the developer to improve the UQAL Portal in the future.

Acknowledgment

We want to thank all respondents that participate in this study. The study was funded by ZG-2019-005 research grant.

References

[1] W. Quan, "Research on development and application of User Experience," Francis Academic Press, UK, pp. 329–332, [2nd International Conference on Mechatronics and Information Technology Research, ICMIT, UK, 2017].

[2] P. Bogaards, and R. Priester, "User Experience:back to business," in Interactions, vol. 12(3), May 2005, pp.23-25.

[3] M. Glanznig, "User experience research: Modelling and describing the subjective," in Interdisciplinary Description of Complex Systems: INDECS, vol. 10(3), 2012, pp.235-247.

[4] M. Hassenzahl, S. Diefenbach, and A. Göritz, "Needs, affect, and interactive products–Facets of user experience," In Interacting with computers, vol. 22(5), 2010, pp.353-362.

[5] A.P. Vermeeren, E.L.C. Law, V. Roto, M. Obrist, J. Hoonhout, and K. Väänänen-Vainio-Mattila, "User experience evaluation methods: current state and development needs," ACM, NordiCHI 2010, Reykjavik, Iceland, pp. 521-530, October 15–19, 2010, [Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries, Reykjavik, Iceland, 2010].

[6] B. Melançon, A. Micka, A. Scavarda, B. Doherty, B. Somers, J. Rodriguez, K. Negyesi, M. Weitzman, R. Scholten, R. Szrama, and S. Boyer, The definitive guide to Drupal 7. 2011. Apress.

[7] K. Lipp, "User experience beyond usability," Technical Report, LMU-MI-2012-2, pp.13-19, September 2012, [Media Informatics Advanced Seminar 'User Behavior', Germany, 2012].

[8] E.L.C. Law, V. Roto, M. Hassenzahl, A.P. Vermeeren, and J. Kort, "Understanding, scoping and defining user experience: a survey approach" ACM, CHI 2009, Boston, MA, USA, pp. 719-728, April 4–9, 2009, [Proceedings of the SIGCHI conference on human factors in computing systems, Boston, MA, USA, 2009].

[9] A. H. Allam, A. Razak, and C. Hussin, "User Experience : Challenges and opportunities," in Journal of Research and Innovation in Information Systems, 2009, pp. 28–36.

[10] N. Ani, H. Noprisson, and N.M. Ali, "Measuring usability and purchase intention for online travel booking: A case study," in International Review of Applied Sciences and Engineering, vol. 10(2), 2019, pp.165-171.

[11] N.M. Ali, A.F. Smeaton, and H. Lee, "Designing an interface for a digital movie browsing system in the film studies domain," in International Journal of Digital Content Technology and its Application (JDCTA), vol. 5(9), 2011, pp.361-370.

[12] N.M. Ali, and A.F. Smeaton, "Exploring the usage of a video application tool: Experiences in film studies," in Informatics in education, vol. 10(2), 2011, pp.163-181.

[13] W.N.W. Ahmad, and N.M. Ali, "The impact of persuasive technology on user emotional experience and user experience over time," in Journal of Information and Communication Technology, vol. 17(4), 2018, pp.601-628.

[14] A. Türkyilmaz, S. Kantar, M.E. Bulak, and O. Uysal, "User experience design: aesthetics or functionality," Managing Intellectual Capital and Innovation for Sustainable and Inclusive Society: Managing Intellectual Capital and Innovation, Bary, Italy, pp.559-565, May 27-29, 2015, [Joint International Conference, Bary, Italy, 2015].

[15] S. Hellweger, and X. Wang, "What is user experience really: towards a UX conceptual framework," 2015, arXiv preprint arXiv:1503.01850.

[16] N.H. Basri, N.L.M. Noor, W.A.W. Adnan, F.M. Saman, and A.H.A., Baharin, "Conceptualizing and understanding user experience," IEEE, Malaysia, pp. 81-84, August 2016, [4th International Conference on User Science and Engineering (i-USEr), Malaysia, 2016].

[17] E. Law, V. Roto, A.P. Vermeeren, J. Kort, and M. Hassenzahl, "Towards a shared definition of user experience," ACM, CHI'08, Florence, Italy, pp. 2395-2398, April 5-10, 2008, [In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, Florence, Italy, 2008].

[18] L. Luther, V. Tiberius, and A. Brem, "User Experience (UX) in business, management, and psychology: A bibliometric mapping of the current state of research," in Multimodal Technologies and Interaction, vol.4(2), 2020, p.18.

[19] R. Bernhaupt, "User experience evaluation in entertainment," in Evaluating user experience in games, 2010, pp. 3-7, Springer, London.

[20] L.E. Nacke, P. Mirza-Babaei, and A. Drachen, "User experience (ux) research in games," ACM, Glasgow, Scotland UK, pp. 1-4, May 4, 2019, [In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 2019].

[21] V. Nagalingam, and R. Ibrahim, "User experience of educational games: a review of the elements," in Procedia Computer Science, vol. 72, 2015, pp.423-433.

[22] Z. Liu, "User experience in Asia," in Journal of Usability Studies, vol. 9(2), 2014, pp.42-50.

[23] K.B. Korasala, and S.S. Duriseti, "Expanding user experience in India," in Journal of Usability Studies, vol. 10(2), 2015, pp.63-67.

[24] M. Pretorius, J. Hobbs, and T. Fenn, "The user experience landscape of South Africa," SAICSIT '15, Stellenbosch, South Africa, pp. 1-9, September 28-30, 2015, [In Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists, Stellenbosch, South Africa, 2015].

[25] B. Alenljung, R. Andreasson, E.A. Billing, J. Lindblom, and R. Lowe, "User experience of conveying emotions by touch," IEEE, Lisbon, Portugal, pp. 1240-1247, Aug 28 - Sept 1, 2017, [In 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man), Lisbon, Portugal, 2017].

[26] T. Grundgeiger, J. Hurtienne, and O. Happel, "Why and how to approach user experience in safety-critical domains: the example of health care," in Human factors, vol. 63(5), 2021, pp.821-832.

[27] F. Erdős, "Economical aspects of UX design and development," IEEE, Naples, Italy, pp. 211-214, October 2019, [In 2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Naples, Italy, 2019].

[28] J.A. Bargas-Avila, and K. Hornbæk, "Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience," ACM, CHI 2011, Vancouver, BC, Canada, pp. 2689-2698, May 7–12, 2011, [In Proceedings of the SIGCHI conference on human factors in computing systems, Vancouver, BC, Canada, May 7–12, 2011].

[29] C.L.B. Maia, and E.S. Furtado, "A systematic review about user experience evaluation," Lecture Notes in Computer Science, Springer, Cham, vol.9746, pp. 445-455, July 2016 [Marcus, A. (eds) Design, User Experience, and Usability: Design Thinking and Methods (DUXU 2016), Springer, Cham, 2016].

[30] J. Nielsen, Ten usability heuristics, 2005.

[31] A.I.I. Paramitha, G.R. Dantes, and G. Indrawan, "The evaluation of web based academic progress information system using heuristic evaluation and user experience questionnaire (UEQ)," IEEE, Palembang, Indonesia, pp. 1-6, October 2018 [2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, 2018].

Design of a Cloud-Blockchain-based Secure Internet of Things Architecture

Deepti Rani, Nasib Singh Gill, Preeti Gulia
Department of Computer Science and Applications
Maharshi Dayanand University
Rohtak, Haryana
India

***Abstract*—The growing number of Internet of Things (IoT) objects and the operational and security challenges in IoT systems are encouraging researchers to design suitable IoT architecture. Enormous data generated in the IoT environment face several kinds of security and privacy challenges. IoT system generally suffers from several issues like data storage, safety, privacy, integrity, transparency, trust, and single point of failure. IoT environment is emerging with several solutions to resolve these problems. The main objective of this paper is to design a cloud-blockchain-based secure IoT architecture that provides advanced and efficient storage and security solutions to IoT ecosystem. Blockchain technology appears to be a decent choice to resolve such kinds of problems. Blockchain technology uses a hash-based cryptographic technique for information security and integrity. Cloud computing provides advanced storage solutions with several remote services to store, compute and analyze the data. The proposed IoT architecture is based on the integration of cloud and blockchain services, which aim to provide transparent, decentralized, and trustworthy and secure storage solutions. In addition to the standard layers (perception layer, network layer, processing layer, and application layer) the proposed IoT architecture in the present paper includes a service layer, a security layer, and a parallel management and control layer, which focus on the security and management of the entire IoT infrastructure.**

Keywords*—Internet of things; cloud computing; blockchain; iot architecture; security and services

I. INTRODUCTION

Data or information security in Internet of Things (IoT) environment is a significant challenge that is getting complicated due to the exponential expansion of IoT devices, applications, and services. In the last few decades, the idea of IoT has widely extended in human life. It brings more ease, upgrades work efficiency, and promotes the growth of national economy of a country. IoT is one of the most promising technologies of the current century which has largely inter-connected the universal things (devices) using various Internet services. The interconnected IoT devices produce a large amount of data that need to be collected, aggregated, stored, and processed privately and securely [1]. But at the same time, IoT technology brings massive security risks for its data and network system.

Security and privacy are the primary requirements for the successful integration of IoT in the society. Due to the absence of proper security, the growing and extensive IoT technology is exposed to various kinds of security and privacy issues [2]. Valuable data in its original form can be captured illegally by cybercriminals from storage (cloud or media) or while communicating over the network. An advanced security system is needed to secure large amount of data generated in IoT system. Data encryption is a worthy approach that provides secure data preservation and data transmission. Blockchain technology is an innovative approach using which data can be securely preserved and transmitted on a decentralized network. Data stored in blockchain is encrypted using secure hash functions (SHA-256 and Keccak-256) that is almost impossible to tamper with [3]. Data can't be updated or deleted by third party due to immutability and integrity properties of blockchain technology. Data generated by various IoT devices can be stored and processed in network edge (device storage) or remote server itself. However, restricted capacity of IoT objects in terms of storage, computational power, and energy is a also significant challenge. Cloud computing provides scalability, management, simplicity, computation, storage and processing facilities to IoT infrastructure [4].

IoT is a large-scale information system that is generally designed using three logical layers which are the perception (sensing) layer, network layer, and application layer. Several researchers have designed many IoT architectures using these three layers for providing solutions to their respective targeted problems such as security and data processing. The present paper proposes an advanced IoT architecture that promotes the solutions for data storage, security, and management problems in IoT based smart environment.

Cloud computing provides centralized storage solutions to IoT infrastructure. It provides features like good scalability, robustness, elasticity, less cost, and power consumption that improves the performance and efficiency of IoT system. Blockchain, as discussed earlier is an innovative distributed technology that provides immutability, integrity, and security to manipulated data. Combining the blockchain technology with cloud infrastructure facilitates with more promising solutions to IoT environment [5]. Initially, the blockchain technology was designed for the security of the public digital ledger of Bitcoin cryptocurrency used for economic transactions only [6]. The Blockchain hypothesis is based on peer-to-peer network architecture in which transaction is not controlled by a single centralized entity. The transaction is stored and controlled in form of blocks in decentralized

manner and these blocks are accessible to all participants of the blockchain network in a trustworthy manner [7]. Both these technologies have brought a great revolution in communication and information in various technical fields, including IoT.

Technologies and components used in IoT systems may create critical security concerns. So, it is important to protect IoT systems in every dimension of IoT infrastructure. Efficient approaches need to be designed for the security of IoT systems. CIA triad is a widely used security model that consists of three major key components: confidentiality, integrity, and availability. These security features must be targeted while designing IoT architecture and ensuring security in every related application area [8].

The proposed architecture presented in the present paper is motivated by several kinds of issues present in the IoT ecosystem. Most of the existing IoT applications suffer from several kinds of security issues (such as privacy, integrity, and single-point of failure) as well as resource related issues (memory, storage, etc.). An ideal IoT architecture must be designed to get rid of all of these problem. The main contributions of the paper are as follows:

- The paper focus on design of an organized IoT architecture utilizing the features of blockchain technology and cloud infrastructure which provide advanced and efficient security, storage and computational solutions.
- Authors in the present paper proposed a 7-layered (perception, network, transport, processing, service, security, application and parallel management and control) Cloud and Blockchain-enabled secure IoT architecture.

Section II presents a brief survey of related literatures. Section III gives a brief description of cloud computing, blockchain technology, and integrated cloud and blockchain technologies which will be deployed to design an advanced and secure IoT architecture. Section IV presents IoT architecture consisting of a perception (sensing) layer, network/ transport layer, processing layer, service layer, security layer, and a management and control layer. The proposed IoT architecture uses edge and fog services for processing, storage, and computation at the local level. This section also presents prominent cloud and blockchain technologies in an integrated form to provide better and innovative services. Section V presents the analysis of proposed cloud-blockchain-based secure IoT architecture and provides brief discussion of improvements done by proposed IoT architectures over recently proposed existing IoT architectures. Section VI concludes the entire research work presented in the present paper.

II. Related Work

Several researchers have conducted number of researches in the area of IoT which address various aspects of designing and modeling. Several researchers have proposed different IoT architectures on the basis of different concerns related to storage, security, communication and services.

Several IoT architectures have been explored in the present paper which vary in terms of the number of layers, type of layers, and terminologies used for layers. Sethi and Sarangi in [9] presented some basic and traditional IoT architectures which include 3-layers and 5-layers IoT architectures. Three layer architecture is composed of the perception layer, the network layer and the application layer. Five-layered architecture includes two additional layers which are the processing layer and the business layer. Wu et al. proposed a five-layer IoT architecture [10]. Many authors including Gokhale et al. [11] and Muhammad et al. [12] discussed the four-layer IoT architecture. However, 3-layer IoT architecture is very common and comprises three key layers [13-15]. Initially, three-layer architecture was accepted for IoT management systems. Four-layer IoT system architecture comprises the sensing layer, the network layer, the service layer, and the application layer. In this way, many pieces of research have been conducted by various researchers to design more advanced IoT architectures and every new research targeted a specific problem to be solved. Hence, the features and behavior of IoT architecture depend on the targeted problems.

In recent years, some architectures have been designed using several advanced technologies. Cloud computing provides cetral data processing facility. Cloud is a part of the middle layer that lies below to the application layer and above to the perception and the network layer [16]. IoT architecture proposed by Hassan and Eassa [4] was dedicated to smart home systems that was designed using cloud computing, context-awareness and some other building blocks.

In terms of IoT, data is significant digital asset that need to be stored with efficient and reliable security solutions. Some of the most recent IoT architectures focused on decentralized blockchain technology in order to provide security, immutablitity and trust to the data generated by IoT system. Many researchers have discussed about decentralized blockchain approach [17]. Zhou et al. in [18] proposed a blockchain-based secure IoT system that facilitates homomorphic computation and system security. Ullah et al. also discussed blockchain-based approach for providing security to IoT.

III. Cloud Computing and Blockchain

The exponential growth of IoT technology creates massive and heterogeneous network traffic. Along with the information this technology generates lots of security issues. An ideal IoT architecture can efficiently deal with such affairs. It must be able to secure the entire IoT infrastructure throughout its layers and able to manage and control all the activities. The proposed framework utilizes cloud computing and blockchain to promote information security.

A. Cloud Computing

Cloud computing is the latest Internet-based technology that provides on-demand availability of resources such as data storage and computing power without direct management by the user. IoT data, services and incidents can be remotely stored, computed and processed over the Internet using cloud services and can be accessed whenever required. Cloud

computing provides many services such as Platform as a service (PaaS), Infrastructure as a service (IaaS), and Software as a Service (SaaS) with affordable cost, scalability, faster speed, high flexibility, reduced complexity, and low risk [19], [20].

1) IaaS: It facilitates on-demand fundamental computing, networking and storage resources to consumers over the Internet on the basis of their request. It is composed of physical and virtual building blocks that provide the facility of execution of workloads and applications without worrying about storage and computation efficiency with little expenses.

2) PaaS: In this cloud computing system, users are facilitated with hardware, software and infrastructure services for developing, executing and managing applications without any expenditure and complexity. The users need to pay only for some resources they utilize. Cloud service providers like Microsoft Azure, Amazon Web Services (AWS), IBM, Google Cloud offer PaaS services.

3) SaaS: Cloud computing provides on–demand software services to the users without direct installing on the system. The users can remotely access these services over the Internet without complex hardware and software management.

B. Blockchain

Initially, blockchain technology was particularly introduced and adopted for Bitcoin cryptocurrency [6]. But in recent years, it is widely deployed in numerous application areas to keep digital records in decentralized and secure manner. It is a good choice for digital forensics to preserve digital evidences with high security, integrity, authenticity and confidentiality. It is a decentralized ledger technology that is anticipated on the peer-to-peer network [21].

A blockchain is a group of interconnected blocks used to store transactional data or events that are managed by all the participants without requiring a central authority manager. It stores event information in such a way that is virtually impossible to add, modify, or delete by unauthorized users. It allows all participating (authorized) users to generate and validate transactions in a peer-to-peer manner. Cryptography [22] and consensus [23] approaches are most significant components of blockchain technology.

Cryptography ensures the security and privacy of data and participants. Cryptographic hash functions are the most widely used techniques adopted by blockchain technology. The term 'Blockchain' is composed of block and chain where a chain is divided into many blocks. The initial block (genesis block) has no parent block and its value is set to zero. The consensus approach provides trust in an untrustworthy environment. It verifies the integrity and trust of the transactions. Being a decentralized technology, each block (node) in the blockchain network stores a copy of the ledger to protect data from a single point of failure.

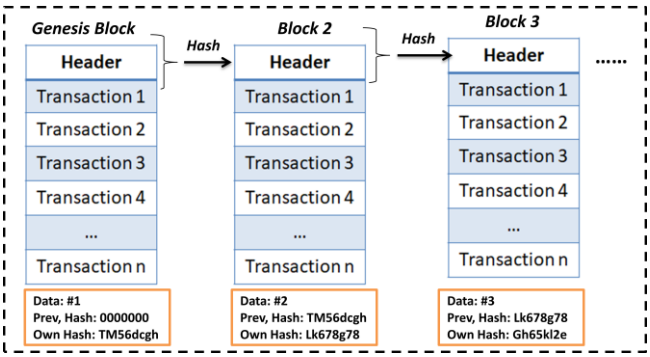


Fig. 1. Linked Blocks.

Fig. 1 shows how blocks are linked and arranged where only the first block has no parent (previous) block to it, hence its previous hash is automatically set to 0 and own hash value is generated for block 1. The hash value of block 1 is passed to block 2 and that value becomes its previous hash then the own hash value of block 2 is generated. The ledger in the blockchain contains a certain number of blocks whereas the first part contains a fact that needs to be stored in a database (e.g., network traffic logs, a record, etc.). The second part contains the header information. It includes the transaction hash, the hash of the previous hash, and the timestamp. This kind of storage makes a sequential chain of blocks [24]. When a new transaction or record needs to be added to a blockchain, first it will be added to a new block. The records added to the blocks in a blockchain can be verified individually using a hash function. Hash functions can ensure data integrity inside blockchain networks [25].

1) Selection of hash functions: A lightweight hash function must be used for block mining. The algorithm used for hashing must serve security using cryptography. The function should also fulfill certain conditions. The output size of a lightweight hash algorithm is 256 bits. If this size is reduced, the security strength also get reduced. The hash algorithms designed for IoT devices need to be designed specifically. These devices do not have sufficient memory of their own to calculate the area for implementation. So some microchips can be embedded in IoT for using cryptographic hash (like SHA-1 and SHA-2) [26].

2) Block structure: Fig. 2 presents the Structure of the Hash Chain or Block. A block structure is composed of a header and the body. The header consists of various fields like Version number, a timestamp, block size, and related transactions [27].

Every transaction generates a hash value to generate a unique Merkle root. Here, the cryptographic nonce is used for proof of the transaction in an encrypted manner i.e. called 'proof-of-work' or 'proof-of-state'. Miner identifies a nonce that generates hash values according to the set value. Difficulty targets the time of block creation. Each block is generated by a distinct hash value. A Merkle tree is a type of binary tree that is formed of hash pointers. It is constructed from leaf nodes toward the head or root node.

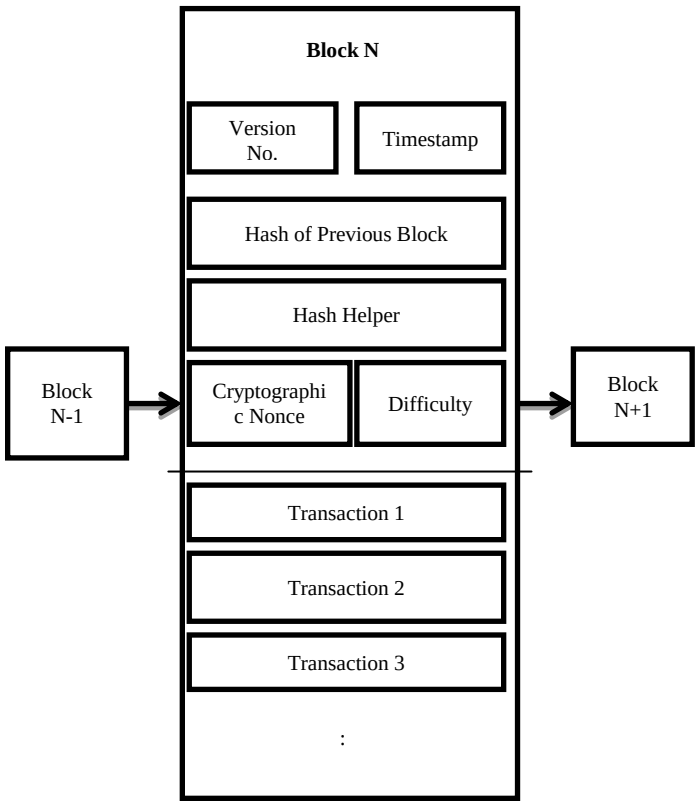


Fig. 2. Blockchain Block Structure [11].

C. Integrated Cloud and Blockchain-Based Internet of Things (BCoT)

The integration of cloud computing in IoT constructs a Cloud of Things (CoT) environment [28]. Cloud computing offers many opportunities in various technical areas including IoT. Cloud computing contributes to making IoT services flexible, scalable, and cost-effective. These services reshape the IoT environment and improve system performance by providing flexibility and robustness. However, the traditional cloud model is based on centralized communication [29]. IoT devices are connected, monitored, and managed by a central server in the cloud. If the centralized system is attacked, it may lead to the destruction of the whole system. An IoT architecture that adopts cloud computing needs a more advanced solution to make the infrastructure decentralized or distributed. It becomes difficult to scale the widespread centralized IoT infrastructure and its communication network requires many communication channels. Several concerns arise in the case of big communication networks like the requirement of a trusted third party (a central cloud server), high communication latency, and increased cost.

To achieve a legitimate solution, a decentralized ecosystem approach needs to be deployed in computing. Blockchain technology is based on peer-to-peer network architecture. The blockchain is a decentralization-based technology and it doesn't depend on the central point of control for transaction management. It also reduces the risk of single-point failure occurring due to disruption of central authority [30]. Each node can be verified independently and the architecture built using blockchain technology ensures secure and robust operations [31]. Integration of blockchain and cloud computing in combined form provides several significant benefits [32]. The blockchain-based architecture uses a distributed storage and computing approach using virtual storage nodes. Blockchain uses a virtual decentralized storage system without using a central authority. Blockchain behaves as a layer among various cloud servers and end-users. Using only cloud services may lead to high-security issues because the centralized server is only responsible for the security and privacy of the entire communication network. Blockchain-as-a-service (BaaS) follows peer-to-peer communication that eliminates the requirement of a trusted third party. Adoption of blockchain technology integrating with cloud computing provides major benefits discussed below:

- Decentralization: It can solve the bottleneck problems occurring in centralized structures, such as single points of failure [33] as well as it also reduces the communication delay and power consumption in IoT devices. It also resolves traffic load balancing issues by establishing short routes [34].
- Enhanced Security and Privacy: For data processing, the Cloud of Things (CoT) need to depend on a third party i.e., a cloud service provider that raises privacy issue. Blockchain-based cloud of Things (BCoT) provides a trustworthy access control that enables only authorized users to access all the services automatically.
- Integrity: Blockchain resolves the problem of data integrity, management and control, and synchronization in distributed databases.
- Quality of Service (QoS): Cloud computing alone is unable to handle Quality of Service for many applications like reliability, real-time, and security. Edge computing is an alternative to cloud computing that can overcome these problems. But it has less scalability, and it is a costly solution. Integrated blockchain with cloud in IoT (BCoT) also resolves this problem [35].
- Immutability: Blockchain provides immutability due to the uniqueness of blocks with unique hash values.
- Scalability: Cloud provides a better storage facility. Integrating blockchain in IoT with the cloud provides better system scalability due to the consensus mechanism.
- Fault tolerance: Replication and redundancy are basic concepts behind fault tolerance which are handled in cloud computing [36].
- Cost optimization: Cloud system provides robust integration of massive data. It also provides distributed resource facility. It improves operational efficiency and reduces cost.
- Strong Authentication: Due to strong encryption and key concepts, a Blockchain-based system provides better authentication.

- Consensus: Consensus is used to establish trust. Consensus may differ in scalability, fault-tolerance, power consumption, etc. [37].

IV. Proposed IoT Architecture

So far, there is no generalized architecture of IoT that has been adopted globally and that can provide various advantages such as efficient storage, decentralized security, and proper data and event management and control altogether in a single IoT architecture. Different researchers have proposed different IoT architectures consisting of different numbers of layers. With the origination of IoT, initially, very simple architectures were proposed that describe the basic scheme of IoT. For many years, three layers architecture consisting of the perception (physical) layer, network layer, and application layer has been widely used. However, this architecture does not provide adequate information about IoT security. It is suitable just for development in the initial stage [10]. It was not sufficient for IoT development; hence a better model was required that can explain the features and inferences of IoT more appropriately.

A. IoT Functional Building Blocks

An ideal IoT framework must consists of all standard components (like sensors, actuators, devices, communication protocols, network and device controllers, etc). The model presented here is composed of seven layers and all the standardized modules (blocks) which are mandatory for an ideal IoT model. IoT system architecture consists of various functional building blocks to assist different utilities viz. sensing, actuation, identification, communication, and management [38], [39].

- Connected Objects (Heterogeneous Things): Internet-connected devices or things are the main components of an IoT system that include sensors, actuators, monitoring devices, Bluetooth Low Energy (BLE) devices, and Radio Frequency Identifiers (RFID). IoT devices are end nodes that can communicate with other connected nodes (devices and applications). These nodes can send and receive data, process the data locally, or get it processed by centralized servers or cloud-based back-ends. All connected nodes generate a certain amount of data in any form that is processed by a data analyzer to generate useful information.
- Communication: This block carries out communication between connected devices and remote servers. It contains components like communication protocols, network enabling devices, etc.
- Processes: Processes are the technologies or functions which are responsible for information processing. The main processes of IoT systems are communication, accumulation, and analysis.
- Services: IoT system performs several types of functions like device modeling, device control, device discovery, data analysis, data control, and data publishing. The service module includes various service-providing technologies such as cloud/fog/edge computing and blockchain technology.
- Management: The management layer contains various functions to govern or monitor activities and components of the entire IoT system.
- Security and Privacy: The security layer provides various functional approaches to secure IoT systems such as authentication, authorization, access control, integrity, privacy, and security.
- Application: The application layer works on the applications of IoT architecture. It functions as an interface for IoT systems and provides the required elements to monitor and control IoT systems. It also allows users to visualize and analyze the present status of the IoT system.

B. Layers of Proposed IoT Architecture

The proposed IoT framework also highlights the concept of the flow of data (information) over a network through an IoT infrastructure. The IoT architecture presented in this paper pays attention to the flow of information through six standard layers that is managed and controlled using a parallel layer. A standard architecture of an IoT system has been depicted in this paper. Fig. 3 presents the proposed IoT architecture enabled with integrated cloud computing and blockchain technology.

1) Perception layer: The perception layer also known as the sensing layer is a physical layer. This layer encompasses many types of sensors. The general idea behind this layer is to collect real-time data from heterogeneous sources such as connected physical and digital devices, controllers, and applications [10]. Sensors are the most important tools which contribute to sensing and collecting essential data from related sources. The sensor senses the presence of a physical thing or a quantity and collects value on a physical parameter e.g., temperature, pulse rate, etc. The sensor returns output in form of signals readable by humans. Transducers are the tools that convert signals from one form to another form. This layer is also known as the physical layer because it can collect data and values directly from devices. The components (Things) of this layer represent the front end of IoT. The ‘things’ are the uniquely identifiable devices carrying a unique IP address that makes them easily identifiable over the network. The perception layer contains various devices such as RFID tags and readers, GPS, cameras, etc. with important sensing technologies.

a) Controller: Microcontrollers are most widely used in IoT technology. A microcontroller operates at the physical/ abstraction layer running the selected operating system/real-time operation system, which provides operating facilities to IoT devices. A microcontroller is a single integrated chip (IC) embedded with a CPU, RAM, ROM memory, and peripherals. It is like a small computer itself. Many things are embedded in a single chip, and it provides lower performance than a microprocessor. But it is a much better choice for smart IoT devices, and its computing power is also sufficient for all IoT applications.

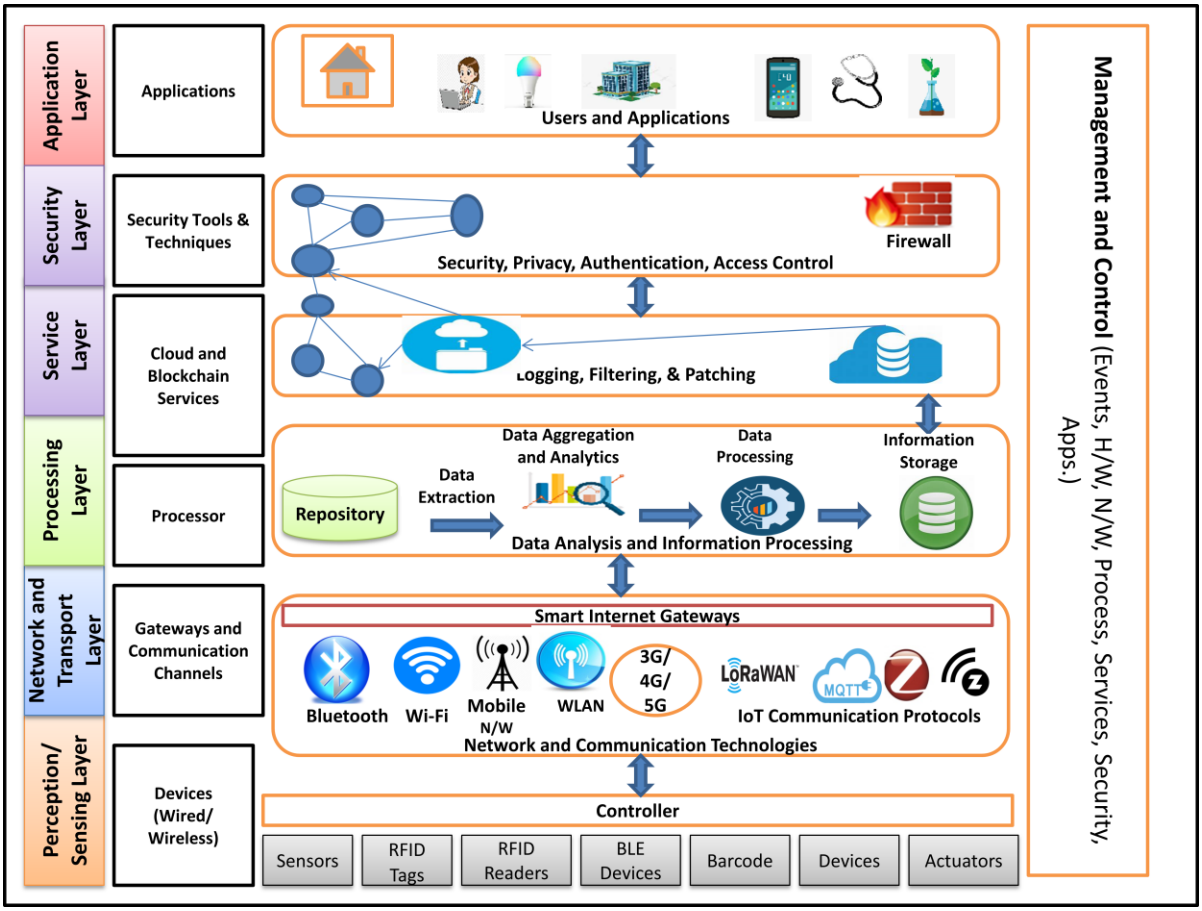


Fig. 3. Blockchain-Enabled Secure IoT Architecture.

2) Network and transport layer: The transport layer receives data packets from the perception layer and transmits them to further layers using various network technologies viz. Bluetooth, Wi-Fi, wireless LAN, and various IoT communication protocols (HTTP, Zigbee, MQTT, Z-Wave, 6LoWPAN, LoRaWAN, etc.) [40]. This layer performs data management from storage to processing state of data leading to reliable data transportation [41].

a) Smart Gateway: It is not possible to directly connect many sensor networks and IoT. IoT and sensor networks consist of sink nodes and base stations. The role of the gateway is to collect the data from base stations and the sink nodes and generate a multi-hop communication structure. In this structure, multiple nodes develop and spread more widely. More heterogeneous data development requires more processing and comprehensive data analysis from the gateway. This structure also customizes the security according to the requirements of IoT and WSN. Sink nodes can efficiently handle the sensor nodes. Here, the gateway handles the heterogeneous data collected from various kinds of devices and networks.

3) Processing layer: In IoT systems, data is captured from devices and stored in datasets, and processed for further requirements. The processing layer is the brain of the IoT system that contributes to the analysis of event data and information processing. Events captured by different sensor devices are stored in datasets. An enormous amount of raw data is extracted from the data repository that is to be analyzed and processed. The processing layer operates on a real-time system and it is responsible for data security through encryption and decryption approaches.

The information gathered from smart IoT devices or end nodes is sent to the cloud [42]. The information is then processed through any processing technique (which could be AI/ Machine learning) for providing data to the user. A certain level of intelligence is added to the devices which makes the IoT system smart. Devices like sensors and actuators are added to the system which collects information from IoT devices and sends it to the cloud through a communication mechanism (Bluetooth, WiFi, Zigbee, etc.) [43].

4) Service layer: The service layer provides a rich set of functions for communication, processing, and storing. This layer is responsible for managing IoT services. Cloud and edge [44] are important IoT services. These services are required to enable IoT systems to meet the security, scalability, and speed requirements [45].

a) Cloud Computing: Technically, the cloud is not a mandatory part of an IoT system. Data storage and data processing take place remotely in the cloud rather than in the system itself. Most of the users store their data on Google

Drive rather than on personal computers. Google Drive uses Google cloud services. Cloud computing makes the Internet of Things more successful by enabling users to perform computing using the services provided over the Internet [46]. A large amount of data is generated by rapidly growing IoT technology that arises the problem of storing, processing, and accessing the data. The integration of cloud computing into the Internet of Things gives an innovative solution. Uploading and storing of sensory data streams take place on the cloud instead of the local device storage.

b) Edge Computing: In IoT edge, devices and sensors communicate real-time data to a network. Pairing the IoT with edge computing makes data processing much faster. In the absence of cloud computing, edge networks and edge computing approaches can control IoT system units individually. The combination of edge and the cloud provides much better development for IoT. Edge computing also reduces the load of data by aggregating it before uploading it to the cloud. Azure is a widely used IoT edge [47]. It allows for storing, processing, and analyzing large volumes of real-time data locally at the network edge. It saves time and resources to send all data to the cloud. Sensitive information is processed and packaged into several packages and sent securely to the cloud.

c) Fog Computing; Fog computing can be used as an alternative to edge computing. If gateways are not able to handle interoperability and trans-coding, this can be attained through fog computing. A fog network is established between the cloud and gateway network. Fog provides better and refined applications and services by extending conventional cloud computing [47]. Fog computing provides a virtual platform that provides storage, computation, and network services between IoT end nodes and the cloud [48]. Fog can provide better quality functioning to mobile nodes by positioning the proxies and access points according to nodes. Fog provides better communication than gateways. It can also include virtual sensor nodes and virtual sensor networks. By co-locating with the smart gateway, it provides low latency communication, temporary storage, better security, more privacy, and easy preprocessing of smart tasks like facilities [49].

d) Blockchain Services; Blockchain technology can be utilized to store digital information in a public database. Some widely known industries like Amazon and Microsoft Azure offer blockchain services i.e. Blockchain as a service (BaaS). A Group of digital information is stored in form of blocks in a hierarchical form. A unique encrypted code is assigned to each block that distinguishes them from each other. These blocks are generally designed using a hash mechanism combined with special programming techniques [50]. It provides improved security by removing human involvement. It uses Asymmetric key cryptography for transactions. Block of keys (public and private keys) are used for entire transactions. Signatures are validated using a private key and a public key verifies the signatures generated by a private key. The decentralization of blockchain makes it harder to tamper with the stored results. However, blockchain services are utilized only to store transactions. The transactions are verified, hashed, and stored in blocks in form of digital signatures.

5) Security layer: Security is the major requirement of IoT architecture. The IoT security layer takes responsibility for managing the security of various components of IoT across the entire infrastructure. The security layer is essential for the security of all the layers of IoT architecture. It makes the information secure before communicating between external and internal users. This layer collects the processed data from the processing layer and encrypts it before sending it over the network using a strong encryption algorithm or a combination of selected encryption algorithms. The legitimate receiver receives the encrypted information that could not be recognized by the illegitimate user [51]. However, securing only the information is not sufficient to secure an IoT infrastructure. It is required to follow all security measures to protect the IoT environment. Authentication and access control mechanisms are used to manage and improve the security system. Blockchain provides security solutions for storage as well as communication.

a) Device Security: Smart devices around the world can communicate with the services like servers or the cloud using Ethernet or Wi-Fi. But these devices are not well-equipped to manage the security concerns of Internet connectivity. The devices must be activated by security features. Security features embedded with hardware and firmware enable devices to handle security, authentication, encryption, proxies, caching, connection loss, timestamps, etc. Device security can be achieved using the following methods:

- Trusted platform module: IoT enabled chips could be embedded with cryptographic keys for the security of end nodes. Security chips containing the security protocols that can be deployed on the sensor devices and these protocols are called to recognize the security operations. Security operations include mutual authentication, mutual signature verification, etc. The security chips may be connected to the sensors using an SD card [52].
- A secure boot process can prevent unauthorized code from the device.
- Security patches must be updated regularly to protect from malware and threats.

b) Network/Communication Security: A communication network is a medium over which data is transmitted and received. Unsecure communication channels might be liable to serious security risks. The communication layer must be equipped with innovative security solutions. Data communication over the network should be encrypted for a secure connection. Data encryption protects communicating data from unauthorized access and information interception. IoT-centric messaging protocols (AMQP, MQTT) can use Transport Layer Security (TLS) cryptographic protocol for end-to-end data protection. However, firewall also works as an obstacle between a secure and insecure network. A firewall is like a physical security fence that monitors the network and

attempts to block certain types of incoming suspected network traffic to prevent attacks on a private network. It does this by filtering both in and out network traffic. Blockchain is a smart security mechanism that provides end-to-end security to entire IoT architecture including cloud.

c) Service Layer Security: Modern IoT systems adopt cloud-based service solutions, which are vulnerable to privacy violations, and failure on a single point, Denial-of-service attacks [53]. Blockchain is an advanced technology that employs a cryptography approach to guarantee the security of distributed ledgers. It supports many advanced technologies such as hashing, elliptic-curve, and distributed consensus approach. Blockchain combined with services provides a promising solution. Cloud is a central system that can be exploited by attackers. Blockchain provides independent and distributed services utilizing public-key cryptography algorithms. Blockchain-based IoT system also facilitates access control.

d) Application layer Security: The application layer is the most sensitive and wider attack surface. Applications are directly exposed to users. The users could be authenticated or malicious actors. The security of the application layer depends on the type of application and the purpose of the application. Security needs to be customized to the unique situation [54]. Trade-offs accompanying strict security measures might be effective in preventing attacks. Application layer security also relies on the selection of communication protocols such as (HTTP, MQTT, CoAP, etc.) used with the system. Each protocol has its strategy to conduct user authentication. So it is important to be familiar with the pattern of each protocol for security improvements. Message Queuing Transport Protocol (MQTT) is the most widely used Client Server publish/subscribe messaging transport protocol. It is a simple, ideal, and above all lightweight messaging protocol used in end-to-end communication. MQTT can be used for telemetry to receive data from sensors and actuators and can command that remotely using the MQTT client library [55]. It supports various authentication mechanisms and Secure Socket Layer (SSL)/Transport Layer Security (TLS) based encryption for transport protection [56]. Application firewalls can be used to guard the application layer. However, firewalls must build and configure considering the specificity of applications. A highly secure application layer can protect other layers too from security breaches because most of the breaches enter through the application layer. It is important to consider security in the designing of protocols [57].

6) Application layer: The application layer is the top layer of IoT architecture. This layer directly interacts with outside users and delivers application-specific services to the users. All the communication from user to system passes through the application layer [58].

a) Authentication: The security mechanism is integrated with the application layer. A user who wants to access an IoT system first needs to pass through the authentication process. Using an identity identification mechanism, the unauthorized user is prevented to access the system.

b) Risk Assessment: The integration of effective security mechanisms in IoT provides an improved security structure.

c) Intrusion Detection: Several application-specific intrusion detection techniques are used with IoT to find security solutions. All the incoming and outgoing events are monitored and their logs are stored in databases which are analyzed for threat detection. An alarm is triggered on the occurrence of suspicious activity.

7) Management and control layer: IoT is an ecosystem where several heterogeneous devices (things) are connected using the Internet carrying distinct missions and functionalities. IoT management has been a challenging task for researchers [59]. In the network, the devices are connected and recognized with the help of their unique IP addresses. IoT device management is like ant colony management. This is a parallel layer that aims to provision, configuration, administration, monitor, and diagnostics of various assets utilized on IoT platforms. It also plays an important role to detect various challenges faced by connected devices.

a) Device Management and Control (DMC): IoT devices can be managed using various tools and techniques designed for IoT device management. IoT device management is used to maintain the security, connectivity, and efficiency of connected smart devices. Management and control Fundamental requirements of IoT device management are:

- Provisioning and Authentication: Provisioning is the process of registering an IoT device to ensure its reliability of an IoT device and authentication is the process through which only devices with valid credentials (certificate/key) are registered. It is necessary to protect the IoT system from malicious attacks.
- Configuration and Control: This is the process of installing a new device using some settings to enable it for working. But only installation does not ensure its performance, functionality, and security from threats. So while configuring the control settings must be fine-tuned for device maintenance and management.
- Monitoring and Diagnostics: To solve very issues, it is necessary to identify them first. A constant monitoring system provides the continuous logging of a device.
- Updates and Maintenance: The software is required to be updated frequently from the moment of installation for the flawless functionality of a device. Devices can be updated and maintained manually as well as remotely.

b) Network Management and Control (NMC): Network management is the process of operating, monitoring, and controlling an entire network to optimize its efficiency [60]. NMC is a diversified authority that provides various network management tools, techniques, protocols, and processes to the network administrator. Network management and control emphasize on management, monitoring, and control of various network components responsible for device connectivity and data communication. Several types of networks enabling

devices (Bluetooth, router, gateways, switches, cables, etc), communicating protocols, technologies (Wi-Fi, 3G/4G, etc.), and services (Internet) are used for connectivity and communication. Network management and monitoring are developed every year and are launched in periodic seasons. NMC monitors and analyzes the network traffic that might contain normal as well as anomalous traffic patterns. Based on the nature of the traffic patterns, it is routed and controlled.

c) Data/Information management: Data management is the process to aggregate and analyzing overall collected valid data and refining it into information [61]. Large volumes of data are produced by different IoT devices and applications [62]. A perfect data management framework is needed that can efficiently collect, manage, and distribute data and must be compatible with existing software and hardware. Data collected from IoT devices are used for analytical purposes. IoT data is processed, managed, and analyzed locally using edge computing and at a centralized level using cloud computing.

d) Security and event management: End-to-end security management is essential for ensuring the privacy and security of IoT devices, services, information, and applications. It protects IoT systems from various attacks. Data, logs, and event monitoring and analysis make IoT systems enable them to protect themselves from various threats and vulnerabilities. But it is difficult to prevent all security risks so an efficient security event management process is required that could ensure rapid recovery. Real-time insight tools and audit trails provide facilities like monitoring, analytics, and log management which can be utilized to get the root cause of an event. This information can be used in digital forensic investigations as evidence.

V. Comparative Analysis and Discussion

Most of the well-known existing IoT architectures generally composed of three or four layers. Traditional IoT architectures do not focus on the storage efficiency and security properties altogether. IoT architecture proposed in the present paper collectively focused on these necessary properties in order to design an efficient IoT architecture. Integrating cloud computing and blockchain technology into IoT ecosystem can provide endless solutions to various kinds of security and storage issues. Table I briefly presents comparative analysis of proposed cloud-blockchain-based secure IoT architectures with some recently proposed advanced IoT architectures.

Quereshi et al. in [63] proposed a cloud-based IoT architecture to overcome the problems of storage and resources. Das et al. in [64] proposed a smart IoT architecture integrating cloud computing with IoT mechanisms. The approaches proposed in [63] and [64] resolve storage and resource related problems. But the centralized storage and access facilies provided by cloud computing may create data security, integrity, and privacy issues. The data stored on cloud platform can be easily compromised by unauthorized users. Author in [65] presented a white paper with his research work that consists of an IoT architecture. The proposed architecture focused on connectivity, data management and application analytics. Qiu et al. in [66] proposed an IoT architecture dynamic blockchain technology. This architecture facilitates decentralized storage with trust, transparency and security. Hou et al. in [67] proposed an IoT architecture using the Blockchain technology. This architecture provides good performance in terms of security, integrity, authentication, reliability and trust. However, storage and scalability still remain significant issues. Which can be resolved by storing the data on cloud platform integrating with blockchain technology. Sharma et al. in [68] focused on advantages of blockchain technology while designing the IoT architecture.

TABLE I. Comaparative Analysis of Proposed Cloud-Blockchain-Based Secure IoT Architecture with Existing IoT Architectures

Literatures	Research Gap	Improvements by proposed Architecture
Qureshi et al. [63]	Security and integrity issues due to centralized storage	Integration with Blockchain technology provides better security and integrity.
Das et al. [64]	Data protection, integrity, security issues	Integrating with Blockchain technology resolves security and integrity issues.
A. Hakim [65]	Not focused on security and resource problems	Focused on security, integrity and storage solutions
Qiu et al. [66]	Focused on bitcoin; Storage issue	Provides advantages of cloud and blockchain technologies and it is not application-specific
Hou et al. [67]	Lack of scalability	Cloud computing can overcome the scalability issue.
Sharmaet al. [68]	Focused only on advantages of blockchain	Proposed architecture consideres advantages as well as disadvantages of cloud and blockchain

IoT architecture proposed in present paper considers the advantages as well as disadvantages of cloud computing and blockchain technology. There are several disadvantages of both of these technologies beside their advantages that must be fixed while implementing them into IoT ecosystem.

VI. Conclusion and Future Scope

The paper presents a design of integrated cloud and blockchain-based secure IoT architecture to resolve various kinds of data security and storage challenges. The proposed cloud-blockchain-based secure IoT architecture is composed of seven layers. In addition to various layers (perception layer, network layer, processing layer, and application layer) which are very common in existing IoT architectures and generally included in the design of every IoT architecture, the proposed IoT architecture includes three additional layers namely the service layer, the security layer, and the management and control layer. Blockchain technology provides end-to-end security solutions with trust, integrity, reliablility and reduces many types of challenges in IoT infrastructure. The service layer is the key layer of IoT architecture that uses integrated features of cloud computing and blockchain. This approach provides decentralized or distributed services in an IoT environment that overcomes various challenges occuring due to centralized communication. It prevents single points of failure, high communication costs, and the need for a central agent. It also provides several security benefits by using advanced cryptographic mechanisms like hashing to encrypt

data and events in an IoT environment. The management and control layer placed in parallel to the entire architecture contributes to monitoring, managing, and controlling various activities and components throughout the IoT system. A secure and well-managed IoT architecture is the basic requirement of successful IoT technology. It is highly needed for realizing the dream of smart cities. Therefore, the proposed IoT architecture can be of greatly helpful for researchers as well as can be used in industries and other private and government sectors for building smart infrastructures. In future, the proposed IoT architecture can be deployed for developing different applications in IoT infrastructure with high security and efficiency.

REFERENCES

[1] S. Bhardwaj and S. Harit, "SDN-Enabled Secure IoT Architecture Development: A Review", *Inventive Communication and Computational Technologies*, 599-619, 2022. https://doi.org/10.1007/978-981-16-5529-6_47.

[2] M. R. Raza, A. Varol, & W. Hussain, "Blockchain-based IoT: An Overview," In: *2021 9th International Symposium on Digital Forensics and Security (ISDFS)*, June 2021, pp. 1-6, IEEE. https://doi.org/10.1109/ISDFS52919.2021.9486360.

[3] S. Sharma, A. Parihar, and K. Gahlot, "Blockchain-Based IoT Architecture," In: P. Raj, A. K. Dubey, A. Kumar, P. S. Rathore (eds) Blockchain, Artificial Intelligence, and the Internet of Things. EAI/Springer Innovations in Communication and Computing. Springer, Cham, 2022. https://doi.org/10.1007/978-3-030-77637-4_10.

[4] A. Z. Hassan Samah and E. E. Ahmed, "A Proposed Architecture for Smart Home Systems Based on IoT, Context-awareness and Cloud Computing," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 13, No. 6, 2022.

[5] H. Ullah, M. Abu-Tair, A. Ali, K. Rabbani, J. Daniel, J. Rafferty, Z. Lin, P. Morrow, and G. Ducatel, "IoT security using Blockchain," In *Essentials of Blockchain Technology*, Chapter 8, pp. 169-188, Chapman and Hall/CRC, 2019.

[6] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review,"* 21260, 2008.

[7] F. Tschorsch, & B. Scheuermann, "Bitcoin and beyond: A technical survey on decentralized digital currencies," *IEEE Communications Surveys & Tutorials*, Vol. 18, No. 3, pp. 2084-2123, 2016.

[8] I. Butun, M. Almgren, V. Gulisano, & M. Papatriantafilou, "Industrial IoT,". Springer International Publishing, 2020.

[9] P. Sethi and S. R. Sarangi, "Internet of Things: Architectures, Protocols, and Applications," Journal of Electrical and Computer Engineering. Hindawi, Vol. 2017, Article ID 9324035, pp. 1-25. https://doi.org/10.1155/2017/9324035.

[10] M. Wu, T. J. Lu, F. Y. Ling, J. Sun, & H. Y. Du, "Research on the architecture of Internet of Things," In *2010 3rd international conference on advanced computer theory and engineering (ICACTE)* , Aug. 2010, Vol. 5, pp. V5-484. IEEE.

[11] P. Gokhale, O. Bhat, S. Bhat, "Introduction to IOT," International Advanced Research Journal in Science, Engineering and Technology, Vol. *5*, No. 1, pp. 41-44, 2018.

[12] M. U. Farooq, M. Waseem, A. Khairi, & S. Mazhar, "A critical analysis on the security concerns of internet of things (IoT)," International Journal of Computer Applications, Vol. 111, No. 7, pp. 1-6, 2015. https://doi.org/10.5120/19547-1280.

[13] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, & W. Zhao, "A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications,". *IEEE internet of things journal*, Vol. *4*, No. 5, pp. 1125-1142, 2017. https://doi.org/10.1109/JIOT.2017.2683200.

[14] H. A. Khattak, M. A. Shah, S. Khan, I. Ali, & M. Imran, "Perception layer security in Internet of Things," Future Generation Computer Systems, Vol. 100, pp. 144-164, 2019. https://doi.org/10.1016/j.future.2019.04.038.

[15] R. Mahmoud, T. Yousuf, F. Aloul, & I. Zualkernan, "Internet of things (IoT) security: Current status, challenges and prospective measures," In *2015 10th international conference for internet technology and secured transactions (ICITST), Dec. 2015*, pp. 336-341. IEEE. https://doi.org/10.1109/ICITST.2015.7412116.

[16] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): a vision, architectural elements, and future directions," Future Generation Computer Systems, Vol. 29, Issue 7, pp. 1645–1660, 2013. https://doi.org/10.1016/j.future.2013.01.010.

[17] Q. Wang, X. Zhu, Y. Ni, L. Gu, & H. Zhu, "Blockchain for the IoT and industrial IoT: A review," *Internet of Things*, Vol. 10, No. 2, 100081, 2020. https://doi.org/10.1016/j.iot.2019.100081.

[18] J. Zhou, Z. Cao, X. Dong, & A. V. Vasilakos, "Security and privacy for cloud-based IoT: Challenges,". In *IEEE Communications Magazine*, Vol. 55, No. 1, pp. 26-33, 2017. https://doi.org/10.1109/MCOM.2017.1600363CM.

[19] P. Srivastava, & R. Khan, "A review paper on cloud computing," *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. *8*, *No.* 6, pp. 17-20, 2018. https://doi.org/10.23956/ijarcsse.v8i6.711.

[20] M. Humayun, "Role of Emerging IoT Big Data and Cloud Computing for Real Time Application," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 4, 2020, https://doi.org/10.14569/IJACSA.2020.0110466.

[21] W. Dai, C. Dai, K. -K. R. Choo, C. Cui, D. Zou and H. Jin, "SDTE: A Secure Blockchain-Based Data Trading Ecosystem," in *IEEE Transactions on Information Forensics and Security*, Vol. 15, pp. 725-737, 2020, https://doi.org/10.1109/TIFS.2019.2928256.

[22] S. Aggarwal, R. Chaudhary, G. S. Aujla, N. Kumar, K-K. R. Choo, & A. Y. Zomaya, "Blockchain for smart communities: Applications, challenges and opportunities," Journal of Network and Computer Applications, Vol. 144, pp. 13-48, 2019, https://doi.org/10.1016/j.jnca.2019.06.018.

[23] M. Wazid, A. K. Das, S. Shetty, & M. Jo, "A tutorial and future research for building a blockchain-based secure communication scheme for Internet of intelligent things," *IEEE Access*, Vol. 8, pp. 88700-88716, 2020, https://doi.org/10.1109/ACCESS.2020.2992467.

[24] E-H Diallo, O. Dib, K. Al Agha, "A scalable blockchain-based scheme for traffic-related data sharing in VANETs," Blockchain: Research and Applications, Vol. 3, Issue 3, 100087, 2022, https://doi.org/10.1016/j.bcra.2022.100087.

[25] J. Ali, T. Ali, Y. Alsaawy, A. S. Khalid, & S. Musa, "Blockchain-based smart-IoT trust zone measurement architecture," In *Proceedings of the International Conference on Omni-Layer Intelligent Systems* , pp. 152-157, May 2019. https://doi.org/10.1145/3312614.3312646.

[26] B. Seok, J. Park, & J. H. Park, "Blockchain-based smart-IoT trust zone measurement architecture," *Applied Sciences*, Vol. *9*(18), 3740, 2019. http://doi.org/10.3390/app9183740.

[27] H. Guo, X. Yu, "A survey on blockchain technology and its security," Blockchain: Research and Applications, Vol. 3, Issue 2, 100067, 2022, https://doi.org/10.1016/j.bcra.2022.100067.

[28] M. S. Karunarathne, S. A. Jones, S. W. Ekanayake, & P. N. Pathirana, "Remote monitoring system enabling cloud technology upon smart phones and inertial sensors for human kinematics," In *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*, Dec. 2014, pp. 137-142. IEEE, http://doi.org/10.1109/BDCloud.2014.62.

[29] B. Kantarci, & H. T. Mouftah, "Sensing services in cloud-centric Internet of Things: A survey, taxonomy and challenges," In *2015 IEEE International Conference on Communication Workshop (ICCW)*, June 2015, pp. 1865-1870. IEEE. http://doi.org/10.4018/978-1-5225-1832-7.ch022.

[30] M. D. Nguyen, N. T. Chau, S. Jung, and S. Jung "A demonstration of malicious insider attacks inside cloud iaas vendor," International Journal of Information and Education Technology, Vol. *4*, No. 6, pp. 483-486, 2014. http://doi.org/10.7763/IJIET.2014.V4.455.

[31] M. Ma, G. Shi, & F. Li, "Privacy-oriented blockchain-based distributed key management architecture for hierarchical access control in the IoT

scenario,". *IEEE access*, Vol. 7, pp. 34045-34059, 2019, https://doi.org/10.1109/ACCESS.2019.2904042.

[32] J. Zou, D. He, S. Zeadally, N. Kumar, H. Wang, & K. R. Choo, "Integrated Blockchain and Cloud Computing Systems: A Systematic Survey, Solutions, and Challenges," *ACM Computing Surveys (CSUR)*, Vol. 54, Issue 8, pp. 1-36, 2021, https://doi.org/10.1145/3456628.

[33] L. Zhou, L. Wang, Y. Sun, P. Lv, "BeeKeeper: A Blockchain-Based IoT System with Secure Storage and Homomorphic Computation," *IEEE Access*, Vol. 6, pp. 43472–43488, 2018, https://doi.org/10.1109/ACCESS.2018.2847632.

[34] T. Wang, "A Study on the Innovative Use of Blockchain in the Human Resources Service Industry", *Wireless Communications and Mobile Computing*, Vol. 2022, Article ID 7798595, 11 pages, 2022, https://doi.org/10.1155/2022/7798595.

[35] T. Mai, H. Yao, N. Zhang, L. Xu, M. Guizani, & S. Guo, "Cloud mining pool aided blockchain-enabled internet of things: An evolutionary game approach," *IEEE Transactions on Cloud Computing, 2021*. https://doi.org/10.1109/TCC.2021.3110965.

[36] S. Abbas, M. A. Talib, A. Ahmed, F. Khan, S. Ahmad, & D-H. Kim, "Blockchain-based authentication in internet of vehicles: a survey," *Sensors*, Vol. 21, No. 23, p. 7927, 2021, https://doi.org/10.3390/s21237927.

[37] W. Viriyasitavat and D. Hoonsopon, "Blockchain characteristics and consensus in modern business processes," Journal of Industrial Information Integration Vol. 13, pp. 32–39, 2019, https://doi.org/10.1016/j.jii.2018.07.004.

[38] P. P. Ray, "A survey on Internet of Things architectures," *Journal of King Saud University-Computer and Information Sciences*, Vol. 30, No. 3, pp. 291-319, 2018, https://doi.org/10.1016/j.jksuci.2016.10.003.

[39] S. Sebastian, & P. P. Ray, "Development of IoT invasive architecture for complying with health of home," *Proceedings of I3CS, Shillong*, 2015, pp. 79-83.

[40] G. Sharma, S. Vidalis, N. Anand, C. Menon, & S. Kumar, "A Survey on Layer-Wise Security Attacks in IoT: Attacks, Countermeasures, and Open-Issues," *Electronics*, Vol. 10, No. 19, 2365, 2021, https://doi.org/10.3390/electronics10192365.

[41] T. Hardjono and N. Smith, "Cloud-based commissioning of constrained devices using permissioned blockchains," In *Proceedings of the 2nd ACM international workshop on IoT privacy, trust, and security*, May 2016, pp. 29-36, https://doi.org/10.1145/2899007.2899012.

[42] D. Rani and N. S. Gill, "Review of various IoT standards and communication protocols," International Journal of Engineering Research and Technology, Vol. 12, No. 5, pp. 647-657, 2019.

[43] M. El-Hajj, A. Fadlallah, M. Chamoun, & A. Serhrouchni, "A survey of internet of things (IoT) authentication schemes," *Sensors*, Vol. 19, No. 5, p. 1141, 2019, https://doi.org/10.3390/s19051141.

[44] C. Luo, L. Xu, D. Li, & W. Wu, "Edge computing integrated with blockchain technologies," In *Complexity and Approximation*, Vol. 12000, pp. 268-288). Springer, Cham, 2020, https://doi.org/10.1007/978-3-030-41672-0_17.

[45] Y. Li, L. Zhu, M. Shen, F. Gao, B. Zheng, X. Du, S. Liu & S. Yin, "CloudShare: Towards a cost-efficient and privacy-preserving alliance cloud using permissioned blockchains," In *International Conference on Mobile Networks and Management*, Springer, Cham, 2017, pp. 339-352. https://doi.org/10.1007/978-3-319-90775-8_27.

[46] O. Ali, M. K. Ishak, M. K. L. Bhatti, I. Khan, & K. I. Kim, "A Comprehensive Review of Internet of Things: Technology Stack, Middlewares, and Fog/Edge Computing Interface," *Sensors*, Vol. 22, No. 3, p. 995, 2022, https://doi.org/10.3390/s22030995.

[47] M. Aazam, & E. N. Huh, "Fog computing and smart gateway based communication for cloud of things," In *2014 International conference on future internet of things and cloud*, Aug. 2014, pp. 464-470. IEEE, https://doi.org/10.1109/FiCloud.2014.83.

[48] F. Bonomi, R. Milito, J. Zhu, & S. Addepalli, "Fog computing and its role in the internet of things," In *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, Aug. 2012, pp. 13-16, https://doi.org/10.1145/2342509.2342513.

[49] S. S. Sarmah, "Application of Block chain in Cloud Computing," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, Vol. 8, No. 12, pp. 4698-4704, 2019. http://doi.org/10.35940/ijitee.L3585.1081219.

[50] M. Burhan, R. A. Rehman, B. Khan, & B. S. Kim, "IoT elements, layered architectures and security issues: A comprehensive survey," *Sensors*, Vol. 18, No. 9, p. 2796, 2018, https://doi.org/10.3390/s18092796.

[51] Rashmi, "IoT (Internet of Things) Concept and Improved Layered Architecture," International Journal of Engineering Development and Research (IJEDR), Vol. 6, No. 2, pp. 481-484.

[52] F. Li, Y. Shi, A. Shinde, J. Ye, & W. Song, "Enhanced cyber-physical security in internet of things through energy auditing,". *IEEE Internet of Things Journal*, Vol. 6, No. 3, pp. 5224-5231, 2019, https://doi.org/10.1109/JIOT.2019.2899492.

[53] H. Zhang, & L. Zhu, "Internet of Things: Key technology, architecture and challenging problems," In *2011 IEEE International Conference on Computer Science and Automation Engineering* , Vol. 4, pp. 507-512, June 2011, IEEE, https://doi.org/10.1109/CSAE.2011.5952899.

[54] R. Ratra, P. Gulia, N. S. Gill, "Evaluation of Re-identification Risk using Anonymization and Differential Privacy in Healthcare," International Journal of Advanced Computer Science and Applications, Vol. 13, No. 2, 2022, https://doi.org/10.14569/IJACSA.2022.0130266.

[55] S. Madakam, , R. Ramaswamy, S. Tripathi, "Internet of Things (IoT): A literature review," *Journal of Computer and Communications*, Vol. 3, pp. 164-173, 2015. http://doi.org/10.4236/jcc.2015.35021.

[56] D. Thangavel, X. Ma, A. Valera, H. X. Tan, & C. K. Y Tan, "Performance evaluation of MQTT and CoAP via a common middleware," In *2014 IEEE ninth international conference on intelligent sensors, sensor networks and information processing (ISSNIP), Apr. 2014*, pp. 1-6. IEEE, http://doi.org/10.1109/ISSNIP.2014.6827678.

[57] E. Rescorla, "The transport layer security (TLS) protocol*," version 1.3* (No. rfc8446), Aug. 2018.

[58] H. Zhang, & L. Zhu, "Internet of Things: Key technology, architecture and challenging problems," In *2011 IEEE International Conference on Computer Science and Automation Engineering, June 2011*, Vol. 4, pp. 507-512. IEEE, http://doi.org/10.1109/CSAE.2011.5952899.

[59] A. R. H. Hussein, "Internet of things (IOT): Research challenges and future applications," *International Journal of Advanced Computer Science and Applications*, Vol. 10, No. 6, pp. 77-82, 2019. http://doi.org/10.14569/IJACSA.2019.0100611.

[60] M. Aboubakar, M. Kellil, & P. Roux, "A review of IoT network management: Current status and perspectives," *Journal of King Saud University-Computer and Information Sciences*, Vol. 34, Issue 7, pp. 4163-4176, 2022. https://doi.org/10.1016/j.jksuci.2021.03.006.

[61] M. Abu-Elkheir, M. Hayajneh, & N. A. Ali, "Data management for the internet of things: Design primitives and solution," *Sensors*, Vol. 13, No. 11, pp. 15582-15612, 2013 https://doi.org/10.3390/s131115582.

[62] A. Chahal, P. Gulia, N. S. Gill, "Different analytical frameworks and bigdata model for Internet of Things," Indonesian Journal of Electrical Engineering and Computer Science Vol. 25, No. 2, Feb. 2022, pp. 1159-1166, https://doi.org/10.11591/ijeecs.v25.i2.pp1159-1166.

[63] Z. Qureshi, N. Agrawal, & D. Chouhan, "Cloud based IOT: Architecture, application, challenges and future," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol. 3, No. 7, pp. 359-368, 2018.

[64] D. Das, S. Banerjee, U. Biswas, "Cloud-Based Smart IoT Architecture and Various Application Domains," In: Al-Turjman, F. (eds) Trends in Cloud-based IoT. EAI/Springer Innovations in Communication and Computing. Springer, Cham, 2020, https://doi.org/10.1007/978-3-030-40037-8_11.

[65] A. El Hakim, "Internet of Things (IoT) System Architecture and Technologies," (IoT) System Architecture and Technologies, White Paper., v1.0, pp. 1-6, 2018 https://doi.org/10.13140/RG.2.2.17046.19521.

[66] H. Qiu, M. Qiu, G. Memmi, Z. Ming, M. Liu, "A Dynamic Scalable Blockchain Based Communication Architecture for IoT," In: Qiu, M. (eds) Smart Blockchain. SmartBlock 2018. Lecture Notes in Computer

Science(), Vol. 11373. Springer, Cham, 2022, https://doi.org/10.1007/978-3-030-05764-0_17.

[67] M. Hou, T. Kang and L. Guo, "A Blockchain Based Architecture for IoT Data Sharing Systems," *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2020, pp. 1-6, https://doi.org/10.1109/PerComWorkshops48775.2020.9156107.

[68] S. Sharma, A. Parihar, K. Gahlot, "Blockchain-Based IoT Architecture," In: Raj, P., Dubey, A.K., Kumar, A., Rathore, P.S. (eds) Blockchain, Artificial Intelligence, and the Internet of Things. EAI/Springer Innovations in Communication and Computing. Springer, Cham, 2022, https://doi.org/10.1007/978-3-030-77637-4_10.

Medical Big Data Analysis using Binary Moth-Flame with Whale Optimization Approach

Saka Uma Maheswara Rao[1]
Department of Computer Science and Systems Engineering
Andhra University College of Engineering (A), Andhra University
Visakhapatnam, India

K Venkata Rao[2]
Department of Computer Science and Systems Engineering
Andhra University College of Engineering (A), Andhra University
Visakhapatnam, India

Prasad Reddy PVGD[3]
Department of Computer Science and Systems Engineering
Andhra University College of Engineering (A), Andhra University
Visakhapatnam, India

***Abstract*—The accurate analysis of medical data is dependent on early disease detection and the value of accuracy is reduced when the medical data quality is poor. However, existing techniques have lower efficiency in handling heterogeneous medical data and the complexity of the features was not enhanced using an optimal feature selection model. The present research work has used the machine learning algorithm effectively for chronic disease prediction such as heart disease, cancer, diabetes, stroke, and arthritis for the frequent communities. The detailed information about the attributes is required to be known as it is significant in analyzing the medical data. The process of selecting the attributes plays an important role in decision-making for medical disease analysis. This research proposes Binary Moth-Flame Optimization (B-MFO) for effective feature selection to achieve higher performance in small and medium datasets. Additionally, the Whale Optimization Algorithm (WOA) is used that showed better performances for LSTM that suited well for the process of classification to predict the time series. The present research work utilizes Spark Streaming layers for data streaming to diagnose using Long Short Term Memory (LSTM) with whale optimization approach which is from the heterogeneous medical data. The proposed B-MFO-WOA method results showed that the proposed method obtained 97.45% accuracy better compared to the existing Modified adaptive neuro-fuzzy inference system of 95.91% of accuracy and B-MFO of 92.43 % accuracy for the models.**

Keywords*—Binary moth-flame optimization; complexity of the features; medical data; long short term memory; spark streaming layers; whale optimization algorithm

I. INTRODUCTION

The healthcare business uses huge data of medical treatment and recordings of every patient. The medical information is recorded and printed with various versions that are converted to digital versions efficiently [1]. Due to the extreme volume of patient details, it is possible for enhancing the quality of health care effectively for saving expenses. All the information present could be used in the multi-health care discipline such that health care illness and surveillance are preventative for management [2]. The Big Data (BD) tools are combined with the machine learning and data mining techniques that showed challenges in areas such as health care, education, transportation, and social media with other networks. The machine learning techniques were used that encompassed the phrase that referred to the large datasets. The regular utilization is progressively performed with little that indicates the basic intricacy and lay the first stone with subsequent ethical and misunderstandings of possible ways [3].

The BD movement is applied to unlock endeavor of large dataset values for making the decision, improving the outcomes, efficiency, and data owners have shown the deliverables. The goals are required to be accomplished, that is, collected, stored, access, managing data with various forms turns the volume with the simple steps [4–7]. Intelligence is applied on data points of high information to improve efficiency. Digital equipment usage increases for model and amount of data increases with unparalleled rate [8–10]. The deep knowledge discovery is computed in big healthcare data which has achieved the best results. However, the selection of the optimum subset of relevant and effective features is used for constructing an accurate model. Thus, the selection of features from a vector of one or zero is used for constructing an accurate model. Apache Spark was deployed in the cloud as it focused to apply on the ML models.

The contribution of the research work is as follows:

- The use of the Whale Optimization Algorithm (WOA) showed better performances for LSTM that suited well for the process of classification to predict the time series and solve optimization problems. To compute simulation of prey search, and prey encircling, humpback whales of bubble-net foraging are mimicked.
- Transfer functions such as U-shaped, V-shaped and S-shaped were applied to convert the continuous value to binary values for the feature selection process using the B-MFO technique.

The organization of the research paper is shown as follows: Section II is the literature review of the existing models and Section III illustrates the proposed method. Section IV shows results and a discussion of the proposed method. The conclusion and future work of the proposed research is given in Section V.

II. Literature Review

Nadimi-Shahraki et al. [11] developed a feature selection technique of B-MFO for the HER medical dataset. The developed model reduced the algorithm performances and therefore, the present research used a binary moth-flame optimization (B-MFO) for the selection of effective features based on the large medical datasets. The features such as S-shaped, U-shaped, and V-shaped transfer functions were used which converted continuous to binary values. The B-MFO technique was applied to use a U-shaped transfer function for feature selection to improve performance in a large dataset. While considering the other datasets (Pima and Lymphography), the suggested B-MFO achieved less accuracy when compared with existing Binary Particle Swarm Optimization (BPSO) method.

Li et al. [12] utilized an optimization approach for reinforcement learning on the Electronic Health Records (EHR) for the treatment. Reinforcement learning provided an efficient path for providing a decision sequentially. The developed model used reinforcement learning for optimizing the treatment for analyzing the diseases, diabetes, and sepsis, and showed complications. The EHRs data was modeled in an environment that obtained a probability that was used for the RL process. The agents were explored better as the basic model was cooperative for multi-agent reinforcement based on the value decomposition. However, the recommended model was additionally required to be extended through the decomposition model and thus the results obtained were better than the existing benchmark models.

Sousa et al. [13] applied the decision-making technique for big data analysis in healthcare organizations and People management. The decision-making process is based on healthcare on big data analysis to support healthcare decisions and applied some techniques to increase efficiency. The suggested model has the limitation of irrelevant feature selection that degrades the performance of the classification accuracy and showed diversity in terms of performance.

Chelladurai and Pandian [14] developed the blockchain based EHR model for an automation system for healthcare. The model could access the health data from one provider to another which remained a challenge when they accessed the health records. The fragmented model launched with the health models was immutable with the patient log by using the modified Merkle tree data to secure the storage. The health records were updated by exchanging information among distinct providers. Even though, the viewership contracts were developed on peer-to-peer blockchain networks and blockchain using Merkle tree generation and hashing that required an extension to ensure the integrity of the content.

Vidhya and Shanmugalakshmi [15] developed a Modified adaptive neuro-fuzzy inference system (M-ANFIS) to analyze the multi-disease using the Big Data (BD) from health care. The health care domain obtained an influence based on the BD that affects the data sources as they are concerned with healthcare organization as it is famous with the volume, complexity, high dynamism, and heterogeneity. The BD analytical techniques utilize the functions, tools, and platforms for realizing it among distinct domains that were affected by various health organizations. The healthcare applications show possible propitious research directions. The multiple diseases were analyzed by using Modified Adaptive Neuro-Fuzzy Inference System (M-ANFIS). Yet, the increasing of sources like audio, video, image, GPS, and medical sensors are having prioritization and designation for the level of patients at the emergency.

Ahmad et al. [16] developed a hybrid ML model for the prediction of mortality in paralytic ileus patients based on EHR. Various machine learning techniques were used including Support Vector Machine with Radial Basis Function (SVM-RBF) for the classification to find the highest rank order among the extracted features. Yet, the developed model required robust models for improving the accuracy of the model to improve the model's feasibility.

Shi [17] developed a novel hybrid deep learning model architecture for the prediction of acute kidney injury based on the patient's record data that included Ultrasound kidney images. The developed model used Convolutional neural networks (CNN) that has Resnet and VGG was made as a hybrid model. The feature maps were concatenated with both types of models for creating the input. However, the suggested model required a continuous optimized approach using the larger clinical database for the paired datasets.

From the literature works, the major problems with big data analytics are the size of the data sets and the complexity with validating long-term predictions for medical diagnostics and treatment. Both the amount of data used in healthcare organizations and the number of data sources are expanding. Healthcare facilities face issues including inconsistent and inaccuracy in patient data as a result of the high speed and growing size of big data. It also has trouble in organizing the data after extracting and integrating them, and more attention is needed to increase accuracy and reduce errors in clinical judgments and other medical tools Therefore, this paper proposed a Binary Moth-Flame with Whale Optimization technique to deal with such issues and the difficulties of implementing big data analytics for the enhancement of healthcare services.

III. Proposed Method

The proposed approach is developed by accumulating enormous amounts of data related to patient care over time in order to comprehend and predict diseases that demands an aggregated approach. While the structured and unstructured data originating from large data sets are collected from clinical and nonclinical modalities to gain information about the disease states. As a result, this study attempts to assess the value of predictive analytics in the health care system by examining the accuracy and other metrics in the provision of medical care.

The block diagram of the proposed Modified feature selection optimization approach is shown in Fig. 1. The block diagram consists of a health care data block which is undergone the process of pre-processing. The pre-processed data is undergone for the feature selection of the data and obtained the outputs. Initially, the collected data that consisted of laboratory reports, imaging reports, medication reports,

medication, caregiver notes, and mortality for both out and in the hospital, etc. are applied for the Spark Streaming layers. The main aim of using the spark streaming layers is to improve the diagnosis of the disease based on the data streaming. However, the spark streaming model required effective features for the model construction. Therefore, an optimum model was used for selecting the relevant subsets that were required for the model, and thus it helps to process further diagnosing the disease using the data streaming layers. The proposed B-MFO-WOA algorithm is generated by developing an initial solution. The image data is pre-processed and the parameters are optimally selected by using the Whale optimization algorithm. For finding the best solution during the exploration and exploitation phase, the B-MFO algorithm is used for the selection of subset features. The three transfer functions such as S, V, and U are integrated for solving the optimization problem. The S, V and U transfer functions select the best values based on the generation of slopes and the saturation. These integrated functions are required for improving the B-MFO algorithm performances. The transfer functions such as S and V helps to convert the continuous variables to a binary value of 0 and 1. The U-shaped transfer function maps the velocity function that is continuously generated with the probability values and updates the particle positions. The integration of all these three transfer functions into the B-MFO algorithm showed an improvement in the performance of searching it on the binary searching space. The search space values are having the probability value for uploading the particle position. Once finding the best features, they are fed to the LSTM classifier that uses a softmax layer and obtains an output as a final label to predict the named data labels. The score for the label is named based on the weighted average for the probability prediction when the classifier is applied to the data. The probability of tagging is done for the labels as 1 and the weight parameter is either set as 0 or 1 for knowing the relative importance of the classifier compared with others. Thus, the health care diagnosis is performed for the data.

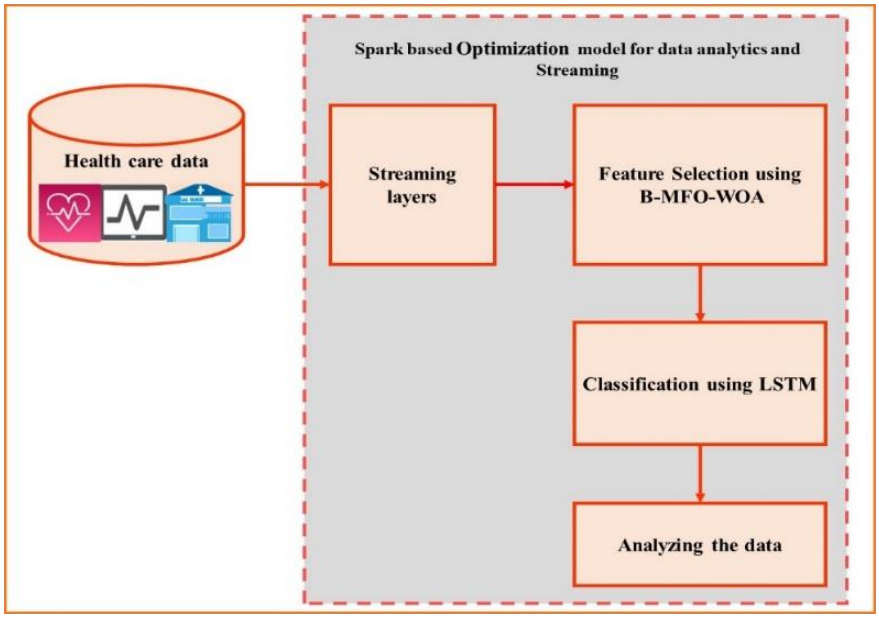


Fig. 1. The B-MFO-WOA and LSTM Model for Medical Data Analysis.

A. Dataset Collection

The health data of the patients are recorded with EHR provides the services for health care in the medical center. The medical centers are registered in detail regarding the patients. The network administrators are registered with the medical centers for participation. The EHR identifies and generates each of the data that is stored in the medical centers. The data is associated total 40,000 number of patients who have stayed in the care units. The units range from the years 2001 and 2012 acquired by Beth Israel Deaconess Medical Center [18]. Initially, the collected data that consisted of laboratory reports, imaging reports, medication reports etc., which are the medical data present in the dataset are applied for the Spark Streaming layers.

B. Feature Selection using B-MFO-WOA

The main aim of using the spark streaming layers is to improve the further diagnosis of the disease based on the data streaming. The stream layers contain polyline, polygon, and point-based features and unlike other feature layers with the services, the data sources have made explicit calls to the data and response for broadcasting the data. In most cases, the features at the irregular intervals are broadcasted and the present research work uses Spark Streaming layers for data streaming to diagnose further.

1) Problem of Feature Selection: The feature process is to select the optimal subset of features to increase the efficiency and relevant features of the model. Accurate data model is constructed and formulated using a feature selection process with a vector having 1 or 0 subsets of features based on the transfer function. It obtains the probability values that change the vector elements that is represented as 0 which can be non-selected and 1 is the selected ones. The feature vector length is the same as the dimensions of the dataset which is used for determining the fitness function which evaluates the subset of

features. This technique reduces the number of features and increases the accuracy of the model. The objective is represented as a fitness function as the CE has shown the error in the classification. N_{sf} and N_{tf} are known as the selected features that are present as a total in the feature dataset. The classification quality significance is given using η and λ (1 − η), as shown in the (1):

$$Fitness = \eta \cdot CE + \lambda \frac{N_{sf}}{N_{tf}} \quad (1)$$

Therefore, the present research uses WOA and B-MFO algorithms for the selection of features.

2) WOA: The WOA uses the exploitation phase for bubble net attacking to perform modeling the bubble net behavior that is having humpback. The two kinds of approaches are designed as follows:

The Shrinking encircling mechanism is performed that achieves the behavior by decreasing the value. The proposed model uses the value to achieve by decreasing a value that represents the fluctuation range when it is decreased. The bubble-net method uses humpback whales to randomly search for prey. Next, in the exploration phase where the prey search is based on the variation in the vector which is used for prey search called exploration. The humpback whales randomly search their positions as per each of their positions. The mutation and evolutionary operations have been included in WOA for formulating and reproducing the behavior of humpback whales that were decided for minimizing the internal parameters and heuristics. This was implemented by the basic WOA version algorithm.

Automatic disease detection is performed using the fitness function for achieving a better classification measure which maximizes the accuracy. The positions for the current solution are updated. The prey is encircled with the phase that performs the process of whale hunting which has started encircled prey position. The whale's best position is found and is considered to be the finest whale. The best whale is towards the other whale which moves once the position is updated. The best solution is determined based on the distances among y^{th} whale where the prey shows the best solution. The distance among the y^{th} whale and the prey calculate the best solution which is ranging between [-1, 1].

3) B-MFO: For finding the best solution during the exploration and exploitation phase, the B-MFO algorithm is used for the optimization of subset features. The three transfer functions such as S, V, and U are integrated for solving the optimization problem. The transfer function of V and S shaped techniques are used in the present research work to convert the MFO function into a binary function. These transfer function names were adapted with the multiple alterable parameters which solved the problem of feature selection. Each of the categories has four versions for transferring the functions and twelve versions had introduced 3 categories for the transfer functions. The datasets such as heart disease, cancer, diabetes, stroke, and arthritis are evaluated for the frequent communities. Additionally, the B-MFO is compared with the best results which are known for its binary metaheuristic optimization approach.

4) B-MFO Variants: The transfer function of S-shaped in B-MFO: The S-shaped or sigmoid function is the transfer function used is named as S_2[100]. The model is introduced originally to develop the binary PSO (BPSO).

$$TF_s(v_i^d(t+1) = \frac{1}{\left(1+exp^{-v_i^d(t)}\right)} \quad (2)$$

From (2), $v_i^d(t)$ is the i^{th} search agent's that is operating at a velocity having the dimension d at the t^{th} iteration. The TFs are converting the probability value of velocity to its next position represented as $x_i^d(t+1)$. The expression is obtained based on the velocity probability value. Here, r is a random value which is ranging from 0 and 1 in (3).

$$x_i^d(t+1) = \begin{cases} 0 & if\ r < TF_s\ (v_i^d(t+1)) \\ 1 & If\ r \geq TF_s(v_i^d(t+1)) \end{cases} \quad (3)$$

From the above expressions, the positions of the search agents are computed based on the current and previous positions. The binary metaheuristic algorithm called BPSO and BGSA is used for transferring the functions. It is used for calculating the probability value that can change the position. The applied transfer function updates the position for each search agent that calculates the probability value. Each of the variants is S-shaped transfer function that showed a slope having S-transfer function. Probability value changes to a positive value as increases in the transfer function. A higher probability function is achieved using S-shaped functions. The S4 provides the lowest value which has affected the position and updates the search agents to find the optimum solution.

C. The Transfer Function of V-Shaped in B-MFO

The V-shaped function is a hyperbolic function that is named with V2 for developing BGSA that has the position to update which is shown in (4).

$$TF_v\left(v_i^d(t+1)\right) = \left|tanh(v_i^d(t))\right| \quad (4)$$

Where t is iteration, d is dimension, velocity of i^{th} search agent is denoted as $v_i^d(t)$. S-shaped function differs from the V-shaped function. The new rules of the updated function are given in (5).

$$x_i^d(t+1) = \begin{cases} -(x_i^d(t) & If\ r < TF_v(v_i^d(t+1)) \\ x_i^d(t) & If\ \ r \geq TF_v(v_i^d(t+1) \end{cases} \quad (5)$$

From the above Equations $x_i^d(t)$ has i^{th} search agent having the position and $-x_i^d(t)$ is the complement value of $x_i^d(t)$. The random value range of 0 and 1 is denoted as r. In case the velocity is low then TF_v encourages the search agents for staying in the positions else the velocity is high. Also, $x_i^d(t)$ has three variants having V-shaped function which is represented as V_1 , V_3 and V_4 are introduced. The higher probability is provided by V_1 than V_2, V_3, and V_4 for the same velocity that affects search agent update and finds an optimum solution.

D. The Transfer Function of U-Shaped in B-MFP

The α and β are two control parameters in the Transfer function of U-shaped that define the slope of U-shaped function width. The U-shaped function is given in (6) and (7).

$$TF_u\left(v_i^d(t+1)\right) = \alpha\left|\left(v_i^d(t)\right)^\beta\right| \quad (6)$$

$\alpha = 1, \beta = 1.5,2,3,4$

$$x_i^d(t+1) = \begin{cases} -(x_i^d(t) & If\ r < TF_u(v_i^d(t+1)) \\ x_i^d(t) & If\ \ r \geq TF_u(v_i^d(t+1) \end{cases} \quad (7)$$

Where t is iteration, d is a dimension, velocity of i^{th} search agent is $v_i^d(t)$ and r of the uniform random number is in the range of 0 and 1. The transfer function of the U-shaped is applied with two conditions. The lower and upper bounds are limited by 1 in (8) and (9).

$$lim_{v_i \to \infty}\, U(v_i^d(t) = 1 \quad (8)$$

$$lim_{v_i \to -\infty}\, U(v_i^d(t) = 1 \quad (9)$$

The variants obtained from the U-Shaped Transfer function is named as U_1, U_2, U_3 and U_4 which were used with the control parameters. The initial iterations were explored for the whole search space of important step that was explored with exploitation with the final iterations. The exploitation step is important for finding a better solution.

The random value for the search space is generated as indicated in (10):

$$E(u) = (e_1, e_2, \ldots e_h) \quad (10)$$

From the above equation (7), E is known as the whales' original population, the interconnected layers with the numbers are represented as h for the process of optimization.

E. B-MFO-WOA

Begin
The population of whales are initialized
Each search agent's fitness function is evaluated
X_{best} = searches for the best search agent
while ($t < maximum\ number\ of\ iterations$)
for each of the search agents:
The positions are Updated as $\alpha, A, C, l\ and\ p$
$if(p < 0.5)$:
$if(|A| < 1)$:
The current agent is updated
else:
The random population of moth is initialized and the objective function is calculated
The set of flames from the same moth is created
The positions of the moths are updated
The flame size has to be changed
End
Return with the best solution

F. Classification

The obtained features are now fed for the LSTM to exhibit the performances. Apache Spark was deployed in the cloud as it focused to apply on the Deep learning LSTM model. The prediction is performed for the higher rate of diagnosis which determines the global best function. The hyperparameters are randomly selected and are passed for the LSTM training. At each iteration, the calculation of parameters is performed. The iteration is stopped when the fitness function is matched The output from the LSTM cell is denoted as h_t, c_t is the memory cell value, LSTM cell output from the previous moment is represented as h_{t-1}. The input data for the LSTM cell is represented as x_t operating at the time t. The process of calculating the LSTM unit is explained in the following steps:

LSTM unit calculation process is explained in steps.

$\widetilde{c_t}$ is known as the candidate memory which is calculated and the bias is represented as b_c. The weight matrix is represented as W_c which is as shown in (11).

$$\widetilde{c_t} = tanh(W_c.[h_{t-1}, x_t] + b_c) \quad (11)$$

The input gate i_t is the current input data that updates the memory cell's state value and controls the input gate. The bias is represented as b_i and the weight matrix is represented as W_i, The sigmoid function is denoted as σ which is shown in (12).

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (12)$$

f_t is the forget gate which calculates the memory state value obtained based on the historic data that updates and controls the forget gate. The bias is represented as b_f and the weight matrix is represented as W_f, as given in (13).

$$f_t = \sigma(W_i.[h_{t-1}, x_t] + b_f) \quad (13)$$

The current moment memory cell c_t is evaluated and the value for the last LSTM unit is denoted as c_{t-1}, as given in (14).

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \quad (14)$$

Where '*' denotes the dot product. Input and forget gate control updates the memory cell based on the state value for the last cell and the candidate value.

Where, o_t is known as the output gate which calculates the memory cell state value as the output is controlled by the output gate as shown in (15).. The bias b_0 and the weight matrix is denoted as W_0.

$$o_t = \sigma(W_0.[h_{t-1}, x_t] + b_0) \quad (15)$$

The output h_t for the LSTM cell is calculated as shown in (16).

$$h_t = o_t * tanh(c_t) \quad (16)$$

LSTM model update, reset, read and keep long time information easily based on memory cell and control gates. The classifier uses a softmax layer for obtaining an output at the final labels for detecting the named data labels. The score for the label is named based on the weighted average for the probability prediction of disease when the classifier is applied

to the data. The probability of tagging is done for the labels as 1 and the weight parameter is either set as 0 for non-diseased labels or 1 for the diseased labels that knew the relative importance of the classifier when compared with others.

IV. Results and Discussion

The proposed model is operating with Python API libraries that are interfaced with the Local Server running in Windows PC 10 pro, 16 GB NVIDIA Geo-force GPU with i9 CPU operating at 2.5GHz.

A. Performance Metrics and Evaluation

The proposed method results are evaluated in terms of performance metrics for the optimized LSTM based model with the Whale Optimizing approach. Indication must be used to guide the use of diagnostic tests in health care settings. Unfortunately, many order tests without taking into account the supporting data. Therefore, in this research, Sensitivity and specificity are crucial test accuracy indicators that enable medical professionals to decide whether a diagnostic tool is appropriate. Healthcare professionals should use diagnostic tests with the appropriate level of assurance in the accuracy, specificity, sensitivity, Area Under Curve (AUC) and Receiver Operating Characteristics (ROC). The mathematical expression for the performances is given in (17–21):

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \quad (17)$$

$$Sensitivity\ or\ Recall = \frac{TP}{TP+FN} \times 100 \quad (18)$$

$$Specificity = \frac{TN}{TN+FP} \times 100 \quad (19)$$

$$AUC = y = f(x) \times 100 \quad (20)$$

where $x = a$ and $y = b$

$$ROC = TPR = \frac{TP}{TP+FN} \times 100 \quad (21)$$

From the above Eq. (17-21), TP is known as True Positive, TN is True Negative, TP is True Positive, TN is True Negative. Table I shows the analysis of different algorithms having data size with 5 GB with feature selection.

TABLE I. The Different Algorithms Having Data Size with 5 GB with Feature Selection

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	ROC (%)
CNN	86	85.25	83.11	84.1	82.23
DNN	90	86.24	84.45	87.45	83.12
LSTM	92	90.8	89.21	91.21	88.24
LSTM based Co-learning model	95.4	92.24	91.21	93.45	91.00
Proposed method (B-MFO-WOA)	99.21	95.45	93.48	95.78	96.87

The health analysis was performed on the patients classified as healthy and unhealthy patients. The results inferred that the percentage for each of the patients is analyzed with respect to the healthy patients with the highest percentage. The existing algorithms used for results analysis are Convolution Neural Network (CNN), Deep Neural Network (DNN), Long Short Term Memory (LSTM), LSTM based Co-learning model. The large training data was needed but failed to encode the position and orientation of the object by using the CNN model. The DNN model was hardware-dependent and showed unexplained behavior in the network when the data were fed. Similarly, the LSTMs showed complexity in the model due to large data set training that needed memory to train. Thus, the existing models showed lower values of performance when compared to the proposed method. Table II shows the evaluation of different clusters that are obtained for different diseases. The present research depicts the number of patients with a particular disease carried out with distinct clusters, patients with various diseases. The existing models such as DNN, CNN, LSTM, LSTM based Co-learning model were used for the evaluation of results in terms of accuracy, sensitivity, specificity, AUC, and ROC. The CNN model obtained 84% of accuracy CNN, 82.95% of sensitivity 81.11% of specificity, AUC of 79.25%, and ROC of 83.02%. The DNN model obtained 87% of accuracy, a sensitivity of 84.24%, specificity of 82.45 %, AUC of 80.65%, and ROC of 84.45%. Also, LSTM model obtained 90% of accuracy, 88.8% of sensitivity, 86.21% of specificity, 86.24% of AUC, and ROC of 88.21%. The existing LSTM based Co-learning model obtained 93.4%, 91.04% of sensitivity, 90.11 % of specificity, AUC of 91.78%, and ROC of 90.99%. The proposed method obtained better accuracy of 95.24%, sensitivity of 92.45%, specificity of 90.4%, AUC of 93.45%, 93.25% of ROC. Fig. 2 illustrates the results obtained for the proposed method with feature selection algorithms.

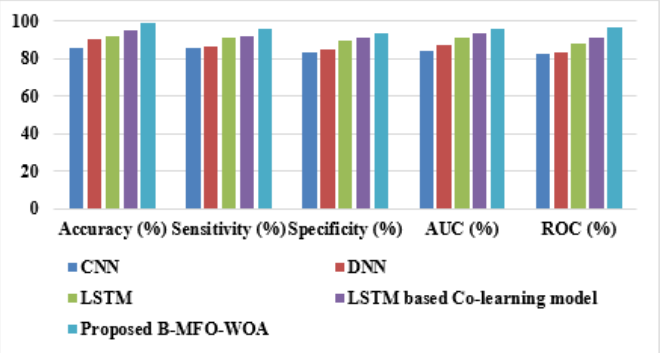


Fig. 2. Results Obtained for the Proposed Method with Feature Selection Algorithm.

TABLE II. Different Algorithms Evaluating Performances for Distinct Data Size with 5 GB Without Feature Selection

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	ROC (%)
CNN	84	82.95	81.11	79.25	83.02
DNN	87	84.24	82.45	80.65	84.45
LSTM	90	88.8	86.21	86.24	88.21
LSTM based Co-learning model	93.4	91.04	90.11	91.78	90.99
Proposed method	95.24	92.45	90.4	93.45	93.25

The existing algorithms used for results analysis are Convolution Neural Network (CNN), Deep Neural Network (DNN), Long Short-Term Memory (LSTM), LSTM based Co-learning model. Table III shows the evaluation of performance metrics for different algorithms having greater than 5 GB data size with feature selection algorithm. From the Table III, it clearly shows that the feature selection algorithm with more than 5 GB data size obtained Accuracy of 88%, Sensitivity of 86.25%, specificity of 90.12%, ROC of 82.02%, AUC of 79.25%. The DNN model obtained 92% of accuracy, 89.24% of sensitivity, 91.45% of specificity, AUC of 80.65%, 83.45% of ROC, LSTM of 93%, 92.8% of Sensitivity, AUC of 86.24%, ROC of 87.21. The LSTM based Co-learning model obtained accuracy of 98.6%, sensitivity of 98.21%, specificity of 97.21%, AUC of 91.75%, and ROC of 90.99%. The proposed method obtained 99.32% of accuracy, sensitivity of 98.98%, specificity of 98.78%, AUC of 95%, ROC of 92.56%.

TABLE III. EVALUATION OF PERFORMANCE METRICS FOR DIFFERENT ALGORITHMS HAVING GREATER THAN 5 GB DATA SIZE WITH FEATURE SELECTION ALGORITHM

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	ROC (%)
CNN	88	86.25	90.12	79.25	82.02
DNN	92	89.24	91.45	80.65	83.45
LSTM	93	92.8	92.8	86.24	87.21
LSTM based Co-learning model	98.6	98.21	97.21	91.78	90.99
Proposed method	99.32	98.98	98.78	95	92.56

Table III show the results obtained by the proposed method that is evaluated using existing algorithms, such as CNN, DNN, LSTM, and LSTM based co-learning model when the data was greater without feature selection and feature selection algorithm. The existing LSTM based Co-learning model obtained 98.6% of accuracy, sensitivity of 98.21%, specificity of 97.21%, AUC of 91.78 %, and ROC of 90.99%. Similarly, the proposed method obtained 99.32 % of accuracy, 98.98 % of Sensitivity, specificity of 98.78%, AUC of 95%, and ROC of 92.56%.

TABLE IV. PERFORMANCE METRICS OBTAINED BY DISTINCT ALGORITHM HAVING DATA SIZE GREATER THAN 5 GB WITHOUT FEATURE SELECTION ALGORITHM

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	ROC (%)
CNN	87	86.14	85.45	85.42	85.24
DNN	91	89.16	88.98	84.57	88.45
LSTM	92	91.23	93.11	86.21	92.21
LSTM based Co-learning model	97.21	94.47	95.211	93.7	96.09
Proposed method	97.45	95.02	96.78	96.87	95.4

Table IV shows the results obtained for different algorithms that are having 5GB greater size without feature selection algorithm. The accuracy of the proposed model without feature selection was obtained as 97.45%, sensitivity of 95.02%, specificity of 96.78%, AUC of 96.87%, and ROC of 95.4%. Fig. 3 illustrates the comparison of results for the proposed method without using the feature selection algorithm.

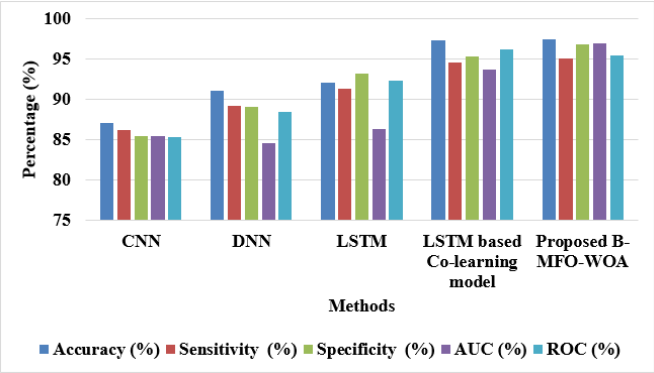


Fig. 3. Comparison of Results for the Proposed Method without using the Feature Selection Algorithm.

B. Comparative Analysis

Table V shows the comparative analysis of the proposed method and existing models evaluated in terms of accuracy, specificity, and sensitivity. The existing B-MFO model needed U-shaped transfer functions for the selection of effective features to overcome the problem of large scale optimization that resulted in Accuracy of 92.43%, Specificity of 92.43%, and sensitivity of 83.51%. Similarly, the Modified adaptive neuro-fuzzy inference system obtained an accuracy of 95.91%, specificity of 98%, and sensitivity of 97.9%. The classification was performed using an SVM-RBF that applied data to the outputs where the classification limited the results for a few of the data values thus obtained an accuracy of 81.30%, sensitivity of 35.59%, and specificity of 91%. The developed CNN model was overloaded with the historic data as it was continuous for data streaming and it was challenging to store, process, and analyse obtained an accuracy of 90% and Sensitivity of 90%. However, the model was required to be extended through the value decomposition model and thus the results obtained were better than the existing benchmark models.

TABLE V. COMPARATIVE ANALYSIS

Method	Dataset	Accuracy (%)	Specificity (%)	Sensitivity (%)
B-MFO [11]	EHR from Beth Israel Deaconess Medical Centre	92.43	96.85	83.51
Modified adaptive neuro-fuzzy inference system [12]		95.91	98	97.9
SVM-RBF [16]		81.30	35.59	91
Convolutional neural networks [17]		90	90	-
Proposed method		97.45	96.78	95.02

V. CONCLUSION

The proposed method showed better performances when operated with B-MFO for the selection of effective features

which were evaluated for large and small medical datasets. The three transfer functions such as S, V, and U-shaped transfer functions are used for the conversion of MFO from the values of continuous to binary values. The WOA is used as an appropriate algorithm to select constrained and unconstrained problems for overcoming the practical applications based on the structural reformation. The combination of the B-MFO-WOA is iteratively executed and is compared with various solutions till an optimum or satisfactory solution is found. The WOA showed better performances for LSTM that suited well for the process of classification to predict the time series. The given time is lagged for an unknown duration of the model as it is based on a deep learning model. The developed model co-learns the best soft labels and deep neural networks based on the training procedure. The Whale optimization approach has the ability for improving the population quality and improves the speed of the algorithm for disease presence prediction. The simulation results showed that the proposed method achieved the objectives by attaining 97.45% accuracy which is better when compared to the existing Modified adaptive neuro-fuzzy inference system of 95.91% of accuracy and B-MFO attained the accuracy of 92.43 %. However, the model showed the complexity in the model due to more features included in a given predictive model which will be analyzed in the future work.

REFERENCES

[1] A. K. Gárate-Escamilla, A. H. E. Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Inf. Med. Unlocked, vol. 19, p. 100330, January 2020.

[2] J. E. Dalton, M. B. Rothberg, N. V. Dawson, N. I. Krieger, D. A. Zidar, and A. T. Perzynski, "Failure of Traditional Risk Factors to Adequately Predict Cardiovascular Events in Older Populations," Journal of the American Geriatrics Society, vol. 68, no. 4, pp. 754–761, April 2020.

[3] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis," IEEE Access, vol. 8, pp. 14659–14674, December 2019.

[4] G. Magesh and P. Swarnalatha, "Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction," Evolutionary Intelligence, vol. 14, no. 2, pp. 583–593, June 2021.

[5] D. Swain, P. Ballal, V. Dolase, B. Dash, and Santhappan, "An Efficient Heart Disease Prediction System Using Machine Learning," in Proc. of ICMLIP 2019, Machine Learning and Information Processing, Advances in Intelligent Systems and Computing, vol. 1101, D. Swain, P. Pattnaik, and P. Gupta, Eds. Singapore: Springer, 2020, pp. 39–50.

[6] S. Sajeev, A. Maeder, S. Champion, A. Beleigoli, C. Ton, X. Kong, and M. Shu, "Deep Learning to Improve Heart Disease Risk Prediction," in Proc. of Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting: 1st International Workshop, MLMECH 2019, H. Liao, S. Balocco, G. Wang et al., Eds. Heidelberg: Springer, 2019, pp. 96–103.

[7] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, "Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique," Journal of Medical Systems, vol. 43, no. 8, p. 272, July 2019.

[8] R. T. Selvi and I. Muthulakshmi, "An optimal artificial neural network based big data application for heart disease diagnosis and classification model," Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 6129–6139, 2021.

[9] H. Das, B. Naik, H. S. Behera, S. Jaiswal, P. Mahato, and M. Rout, "Biomedical data analysis using neuro-fuzzy model with post-feature reduction," Journal of King Saud University-Computer and Information Sciences, 2020.

[10] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.

[11] M. H. Nadimi-Shahraki, M. Banaie-Dezfouli, H. Zamani, S. Taghian, and S. Mirjalili, "B-MFO: a binary moth-flame optimization for feature selection from medical datasets," Computers, vol. 10, no. 11, p. 136, October 2021.

[12] T. Li, Z. Wang, W. Lu, Q. Zhang, and D. Li, "Electronic health records based reinforcement learning for treatment optimizing," Inf. Syst., vol. 104, p. 101878, February 2022.

[13] M. J. Sousa, A. M. Pesqueira, C. Lemos, M. Sousa, and Á. Rocha, "Decision-making based on big data analytics for people management in healthcare organizations," Journal of Medical Systems, vol. 43, no. 9, pp. 1–10, 2019.

[14] U. Chelladurai and S. Pandian, "A novel blockchain based electronic health record automation system for healthcare," J. Ambient Intell. Hum. Comput., vol. 13, no. 1, pp. 693–703, 2022.

[15] K. Vidhya and R. Shanmugalakshmi, "Modified adaptive neuro-fuzzy inference system (M-ANFIS) based multi-disease analysis of healthcare Big Data," The Journal of Supercomputing, vol. 76, no. 11, pp. 8657–8678, 2020.

[16] F. S. Ahmad, L. Ali, H. A. Khattak, T. Hameed, I. Wajahat, S. Kadry, and S. A. C. Bukhari, "A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs)," J. Ambient Intell. Hum. Comput., vol. 12, no. 3, pp. 3283–3293, 2021.

[17] S. Shi, "A novel hybrid deep learning architecture for predicting acute kidney injury using patient record data and ultrasound kidney images," Applied Artificial Intelligence, vol. 35, no. 15, pp. 1329–1345, September 2021.

[18] D. Kalra, "Electronic Health Record Standards," Yearbook of Medical Informatics, vol. 15, no. 01, pp. 136–144, 2006.

Implementation of a Mobile Application based on the Convolutional Neural Network for the Diagnosis of Pneumonia

Jazmin Flores-Rodriguez, Michael Cabanillas-Carbonell
Facultad de Ingeniería, Universidad Privada del Norte
Lima, Perú

***Abstract*—Pneumonia is the main cause of infant mortality in Peru, which has led to plansfig, such as vaccination campaigns, greater economic investment in health, and the strengthening of specialized medical personnel, however, mortality rates remain high. In this sense, the implementation of new computer technologies such as Deep Learning through the use of the artificial neural network is proposed. The objective of this project was to determine the influence of a mobile application based on a Convolutional Neural Network for the diagnosis of Pneumonia, the project consists of the analysis of images of Chest X-rays with Pneumonia and Normal by means of an application developed called "Diagnost". The study was carried out considering a control group and a study group formed by 33 medical staff members who used the application. The analysis of the data obtained was made based on the study of 3 indicators, detection time, result in accuracy, and reduction of medical assistance. According to the results, it was concluded that the mobile application based on the convolutional neural network allows the early detection of Pneumonia and allows the reduction of medical assistance, however, it is still necessary to continue working on the accuracy of the diagnosis.**

Keywords*—Pneumonia; convolutional neural network; deep learning; chest x-rays

I. Introduction

Respiratory infections contribute to high mortality rates worldwide, accounting for more than 4.5 million deaths per year, especially in low- and middle-income countries [1]. These infections cause upper respiratory tract diseases such as rhinitis, sinusitis, pharyngitis, and lower respiratory tract diseases such as bronchitis and pneumonia [2].

The world health organization (WHO) states that Pneumonia is the leading cause of infant mortality in children under 5 years of age worldwide [3], where 2.4 million of these occurred in the first month of life, as well as 1.5 million deaths at the age of 1 to 11 months of life, and 1.3 million at the age of 1 to 4 years [4]. Community-acquired pneumonia is an acute inflammation of the lung parenchyma from microorganisms and is demonstrated by radiological changes in patients and systemic infections, where pneumococcus is probably the most frequent germ. The cause of pneumonia can be fungal, bacterial, or viral, where pneumonia caused by bacteria can be easily treated with antibiotics, but only one-third of children with pneumonia receive the correct medication [5]. According to different studies on the increase of pneumonia in the seasons [6], [7], [8] it has been proved that the belief that low temperatures in winter are the major cause of pneumonia is false; most of the origins of this disease do not depend on seasonal causes.

In Peru in 2021, 1153 cases of pneumonia and 12 deaths from the same cause were reported in children under 5 years of age; so far in the current year 2022, 2149 cases of pneumonia and 21 deaths from the same cause have been reported in the same age range [9], [10]. According to Peru, pneumonia is one of the main causes of death, especially in the pediatric population, highlighting that the method of prevention is timely diagnosis and adequate treatment to help reduce its fatal consequences [11], [12].

The Ministry of Health has carried out contingency plans such as mass vaccination campaigns against pneumonia and influenza, greater economic investment in health, and the reinforcement of specialized medical personnel; however, mortality rates are still high, making it essential to implement new techniques that contribute to the early diagnosis of pneumonia.

Therefore, the aim of this research is to determine the influence of Convolutional Neural Network-based mobile applications on Pneumonia diagnosis. Focusing on early detection of pneumonia, the accuracy of detection, and reduction of medical assistance is in order to contribute to the reduction of pneumonia cases, focusing on early detection by making use of a deep learning technology tool provided to physicians by reducing the demand for these in the detection of pneumonia.

The research is organized as follows. Section II contains a bibliographic study of previous research and its results obtained. Section III details the concepts related to the convolutional neural network for the diagnosis of pneumonia. Section IV formulates the methodology to be used and the type of research. Section V is the case study, where the development and training of the convolutional neural network are described, as well as the development of the application. Section VI is the results phase, showing the descriptive analysis and inferential results. In Section VII, the discussions, analysis, and interpretation of the results are developed. Finally, in Section VIII, conclusions are drawn to enhance the proposed objective.

II. Bibliographic Study

According to [13], a fast and reliable detection method is required to prevent the spread of infections. In recent years, great expectations have been raised from the community of healthcare professionals and patients regarding the use of smart technology to provide innovative solutions for the treatment of diseases this due to the great potential of the technology [14].

In [15] it is stated that regularly the users are not informed about the treatments or the symptoms related to the disease and in case of small problems the user has to go personally for a check-up and which is more time-consuming.

In [16], investigated the case of severe retinopathy of prematurity leading to newborn limitations or blindness, and the need for accuracy and interpretation in deep learning for medical care of these cases, in that study, was used as study material data of premature infants between the years 2011-2014, with the registration of 5000 patients showing 102 variables, where it was necessary to group the records by variants, thus the sample size of 385 patients was determined, and a simple model was built to predict patients with severe retinopathy, the objective of the research is to provide physicians with an interpretable machine learning model. The research that was done allowed us to know that the construction of a deep learning model requires a considerable amount of data and that this model can be an ally to our physicians and that it is easy to understand [17].

The research article [18] on the classification of X-ray chest images using a Convolutional Neural Network aimed to classify chest images previously captured by the cameras of a system, so the convolutional neural network was built, since it has a high performance, being used in problems of classification of images, signals or medical images. For the classification of these images are used large networks previously trained, this network with a real-time application allows the classification to occur in less time. During the training of the network to identify anomalies of the images it is required to classify the data in subsets and increase the data. Likewise, for the present research, the convolutional neural network and the classification of X-ray images will be used.

The research article [19], proposed a method for the classification of pathologies in chest X-ray images based on deep learning or deep learning, for this purpose it makes use of a large number of X-ray images, and from the images, a classification was made between the pathology of pulmonary nodules and cardiomegaly, from that it was concluded that the results obtained showed an improvement for the detection of nodules and cardiomegaly compared to existing methods. In comparison to the present investigation, the use of a deep learning model was made for the classification of X-ray images for pneumonia disease.

In [20] proposed a method for diagnosis based on an imaging study of patients with pneumonia by means of deep learning techniques, in order to achieve a distinction between patients with pneumonia and healthy patients, as well as to differentiate viral between and bacterial pneumonia. The model achieved acceptable results, but with certain limitations in the classification of viral and bacterial pneumonia.

The article [21] proposed a new architecture based on ResNet 50 with some adjustments, for the analysis of medical images of the chest to highlight examples infected with pneumonia. The aim of the research was to highlight the use of machine learning to create a model with accuracy that correctly answers the questions posed to it in the shortest time. The image classification obtained an accuracy of 97.56%. Likewise, in the present research, use was made of the deep neural network trained CVPR 2015 or ResNet152, from this network the weights and parameters necessary for the training of the convolutional network are obtained.

III. Concepts Related to Convolutional Neural Network for Pneumonia Piagnosis

A. Convolutional Neural Network

1) Machine learning: Machine learning is a type of artificial intelligence technique where computers learn to do something without being programmed to do it. The program learns and associates combinations of distinctive features, resulting in a learning process also known as "building a model".

2) Deep learning: Este permite el proceso de Machine Learning por medio del uso de la red neuronal artificial , la cual se realiza por medio de niveles, siendo el primer nivel bastante simple enviando esta información al siguiente nivel, aquí es donde la información sencilla es combinada volviendola más compleja al seguir enviando la información obtenida de forma sucesiva a más niveles.

3) Artificial neural networks: Artificial neural networks are an imitation of the behavior of our neurons in the brain. Our brain has a huge number of neurons connected to each other, forming a neural network, these neurons have three main parts: the dendrites, the body, and the axon. Where the dendrites are in charge of transporting the electronic pulses to the body of the neuron, the body recognizes and works on the signals that arrive and the axon is a single nerve fiber in charge of communicating the body of a neuron with the others. The synapse originates with the contact between the oxon and the dendrite of another neuron [22].

4) Neural network model: A neural network is the joint use of many single neurons, this neural network is made up of hundreds or even thousands of neurons, this is where the concept of layer appears, which is the grouping of all neurons in several sets within the neural network (Fig. 1), where each layer has its own weight matrices, its bias vectors and their respective outputs [22]. Furthermore, the inputs of the subsequent layers are the outputs of the lower layers.

5) Red convolutional: He mentions that a convolutional network is a kind of multilayer network that consists of different alternating convolutional and subsampling layers, finally, it has a sequence of layers that are completely connected like a multilayer perceptron network. In addition, the input of a convolutional network is usually an image of m x m x r, where m is the height as the width of an image and r is the number of channels and it works with grayscale where r=1. Convolutional layers have k filters or Kernels.

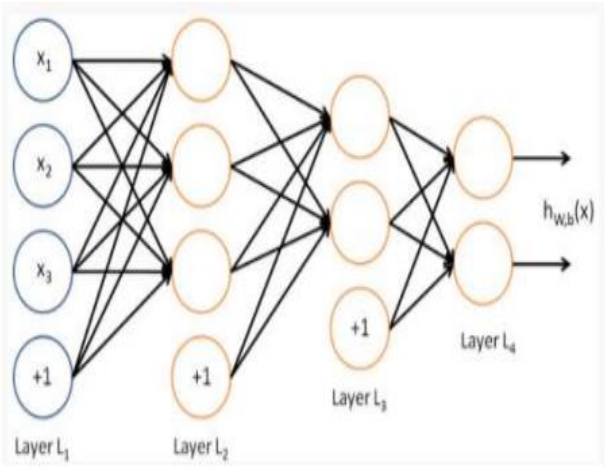


Fig. 1. Multi-layer Neural Network [22].

Layers of a convolutional network:

a) Input layers: The input layer of a convolutional network is an image.

b) Convolutional layer: In this layer, the reduction of the number of possible connections between the neurons of the hidden layer and elements of the input image is carried out, which consists of reducing the computational load of the system. This layer allows extracting useful features from the images to help with their analysis [22]. The convolution is an operation of products and sums between the input image and the Kernel filter, this generates a feature map, where the advantage is that the filter used serves to extract the same feature in any part of the image.

c) Pooling or subsampling layer: In this layer, we use the characteristics of the images obtained in the convolution layer to classify them. The objective of this layer is to support the image characteristics obtained and locate the predominant features of the image. This layer has two types of pooling, pooling or overage-pooling, and max-pooling. In the overage-pooling the elements of the submatrix are selected and their average is calculated and the result is stored in the first position of a matrix which is the output, on the contrary, in a max-pooling, the element with value is searched and this goes to the first position of the output matrix (Fig. 2).

d) Full-Connected Layer: This is the last layer of a convolutional neural network, where we try to classify to determine to which class each input image belongs. In this layer, each neuron is connected with each and every one of the elements of the matrix of the previous layer.

e) Pre-Training. The training of a convolutional neural network is important for the transfer of learning, so it is essential to use pre-trained networks with different applications. These pre-trained models are successfully applied, and the use of these pre-trained networks can be to make small adjustments or function as a feature extractor to achieve a better performance of the data to be processed. Pre-training means initializing the networks with previously trained parameters, instead of setting parameters randomly [23].

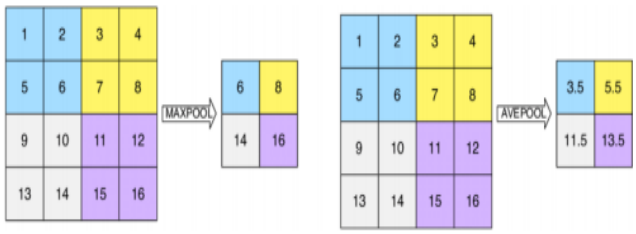


Fig. 2. Differences in Max-pooling and Overage-pooling Operation [22].

6) Respiratory diseases: According to [24], most diseases of the respiratory system fall into one of three categories: obstructive pulmonary diseases, restrictive disorders, and pulmonary diseases.

Within the category of obstructive pulmonary diseases, we have all disorders of the respiratory tract such as asthma, bronchiolitis, etc. Restrictive disorders, also known as parenchymal diseases, include anomalies of the chest wall, as well as neuromuscular diseases. Also [24] states that studies of patients with respiratory diseases begin with a complete anamnesis, this anamnesis should focus on the factors that trigger dyspnea such as cough, which are the cardinal symptoms of respiratory disease.

IV. Methodology

A. Type of Research

The present research is of the applied type since the knowledge acquired by practice is applied in most cases for the benefit of society [25]. The research design is experimental because one or more study variables are manipulated in order to observe the effect of one variable (independent) on another variable (dependent). This is done in order to discover the cause of a particular situation or event [26]. In addition, the type of research is pure experimental type, this type of research meets the requirements to achieve internal validity and control because it has comparison groups. Within this type of research, pretests and posttests can be used to study the evolution before and after the experimental treatment [27].

The independent variable for the research was "Convolutional Neural Network based mobile application", while the dependent variable was "Diagnosis of Pneumonia".

B. Population and Sample

The target population of this research project is the 80 users and/or medical personnel of the hospitals and clinics that will use the mobile application.

The statistical formula (1) was used to determine the sample size.

$$\boldsymbol{n} = \frac{N * Z_{\alpha}^{2} * p * q}{e^{2} * (N-1) + Z_{\alpha}^{2} * p * q} \quad (1)$$

The research population considered a population of 80 members of the hospital's medical staff who will use the mobile application, with a confidence level of 95% and a margin of error of 5%. A 50% probability was considered that the event studied would occur and a 50% probability that the event would not occur. The Table I is the detail of the results of the sample:

TABLE I. SAMPLE VALUES

Nomenclature	Parameter	Value
Population or Universe Size	N	80
The statistical parameter that depends on the Confidence Level (CN)	Z	1.960
(1 - *p*) = Probability that the studied event does not occur.	p	50.00%
(1 – *p*) = Probabilidad de que no ocurra el evento estudiado	q	50.00%
Maximum accepted estimation error	e	3.00%

The sample under study for this research project consisted of 67 members of the hospital's medical staff who will use the mobile application. This was randomly divided between the experimental group (RG1) and the control group (RG2), with 33 members in each group.

C. Indicators

For this research, three indicators were taken into account: "Detection time" (early detection is evaluated); "Outcome" (detection accuracy is evaluated); and "Time to care" (reduction of medical assistance is evaluated).

D. Techniques and Instruments for Data Collection and Analysis

In order to collect data on the variables, it was necessary to collect data by means of detailed procedures using techniques and instruments.

1) Techniques: For the present research study, the technique of observation and data collection was used. Where observation was used to review the evaluation of the influence of the dependent variable on the independent variable.

2) Instruments: Data collection was carried out through the Quantitative Observation Sheet instrument, and then a statistical analysis of the data collected was performed.

3) Procedure: The next step consisted of analyzing the indicators, frequency tables and formulas, through calculations in SPSS software.

E. Methodology for Development

A comparison of agile methodologies was made in Table II in order to choose the one that best suited the proposed project.

For this research, the agile methodology called Scrum was chosen since it will have a short duration and it allows us to have a broad and interactive control of the processes. Scrum is a reference framework within the Agile software development methodology that allows for the creation of complex software and delivers it in a simpler way compared to the waterfall methodology. Scrum proposes short iterative cycles that last about a week or even a month, this period of work is known as iteration or sprint [28].

The framework allows for increased productivity and creativity during the project, enhances team engagement, and enables collaborative task completion [29], as well as understanding the customer's needs as a team to collaborate and deliver maximum value during each iteration.

TABLE II. COMPARISON OF AGILE METHODOLOGIES

Methodologies	Characteristics	Roles	Advantages
RUP	They present reiterative development.	Analysts, Developers, Managers, Stakeholders, specialists, and reviewers.	Extensive documentation, software quality checking, configuration, and change control.
SCRUM	They are used in environments based on agile software development.	Product Owner Scrum Master Developer Teams.	Advantages such as higher project productivity, transparency in the development of processes, and greater control
XP	Continuous, repeated, and automated unit testing.	Programmer, tester, client, follow-up, coach, consultant, and manager.	Communication, simplicity, feedback, decreased trace errors, and high-quality minimum time.

V. CASE STUDY

A. Solution Development

The research developed was technically feasible, since everything necessary for the development of the mobile application and the training of the convolutional neural network is accessible. Tables III and IV below detail the technical aspects considered for the development of the research.

1) User equipment: Regarding the requirements of the mobile equipment on the users' side, the following characteristics specified in Table III were recommended.

TABLE III. CHARACTERISTICS OF MOBILE EQUIPMENT

Characteristics	Optimo
Display	5.1
Battery	300 mAh
Memory	4 GB RAM, 32 GB ROM
Processor	Qualcomm dual core 2.15 GHz+ dual core 1.593 GHz
Operating System	S0 Android 6.0.1
Keyboard	Touch screen with on-screen keyboard
Web	2G/3G/4G/LTE capable

2) Software platforms: During the development of the project, the list of necessary software was divided in Table IV shows the necessary software to be used for the development of the mobile application.

3) Development: For the development of each Sprint, reviews, and deliverables were planned to validate the progress obtained at the end of the Sprints. The estimated speed of the Sprint was 14 days.

a) Sprint N°1: In this phase, the data processing was developed which consisted of Thorax X-Ray images, these were divided into two groups: Training and Test stored in Drive. Likewise, the construction of the neural network and its training were carried out.

TABLE IV. LIST OF SOFTWARE REQUIRED FOR THE DEVELOPMENT OF THE MOBILE APPLICATION

Topic	*Description*
Smartphone (emulator)	Samsung S7
Mobile Operating System	Android 8.0.0
Data storage	SharedPreferences
Programming	Android
Development environment	Android Studio
Play Store	Online store service that allows the distribution of apps
Portátil Asus	Windows 10, intel core i5, eighth generation, Ram 16 GB, hard disk of 1tb.
Operating System	Microsoft Windows 10
Programming	Python
Libraries	Tensorflow, numpy, keras, python, json, panda, matplolib, gdown, sklearn.
Development environment	Google Colab, Keras, Python, Miniconda, Tensor Flow
Dataset	Mendeley Repository
Server	Centos 7, Google Cloud Platform
Pre-trained model	DenseNet
Postman	Perform Get and Post requests
Deep learning technology	Deep learning course on the Udemy platform
Google Drive	Dataset storage and model colab
Google Books	Purchase of books for this research

In Fig. 3, the project folders were defined and the folders of X-ray images of the thorax with Pneumonia and Normal were read.

```
# Definir el Folder principal del proyecto
project_folder = "/content/drive/MyDrive/deep-learning-2/rayosX-tesis-2021"

# Libreria glob: para leer el contenido de cada carpeta
files_train_neumonia = glob.glob(project_folder+"/dataset/train/neumonia/*.jpg")
files_test_neumonia = glob.glob(project_folder+"/dataset/test/neumonia/*.jpg")
files_train_normal = glob.glob(project_folder+"/dataset/train/normal/*.jpg")
files_test_normal = glob.glob(project_folder+"/dataset/test/normal/*.jpg")

# Obteniendo imágenes al azar de cada folder, usamos la libreria image de Keras
# Cargamos las imagenes dentro de la variable: image_...
file_train_neumonia = files_train_neumonia[randrange(len(files_train_neumonia))]
image_train_neumonia = image.load_img(file_train_neumonia)

file_test_neumonia = files_test_neumonia[randrange(len(files_test_neumonia))]
image_test_neumonia = image.load_img(file_test_neumonia)

file_train_normal = files_train_normal[randrange(len(files_train_normal))]
image_train_normal = image.load_img(file_train_normal)
```

Fig. 3. Assigning Image Folders to Variables.

Fig. 4 shows the dimensioning of the established images and Fig. 5 shows the data augmentation using the ImageDataGenerator libraries of Keras:

```
from keras.preprocessing.image import ImageDataGenerator

# Dimensión de las imgs a procesar
img_width = 224
img_height = 224
batch_size = 40
```

Fig. 4. Size the Images to 224 x 224.

```
# Data Augmentation and Normalization (Aumentamos lacantidad de imagenes y las normalizamos dividiendo entre 255)
datagen_train = ImageDataGenerator(rescale=1.0/255.0,      # Normalizar los valores al rango [0-1]
                                   horizontal_flip=True,   # Giro horizontal
                                   rotation_range=15,      # Giro aleatorio (clockwise) entre 0 y 15 grados
                                   width_shift_range=0.15, # Mover la img horizontelmente 15%
                                   height_shift_range=0.15, # Mover la img verticalmente 15%
                                   zoom_range=0.2)         # Zoom in / Zoom out aleatorio de 20% => 80% - 120%
```

Fig. 5. Command for Data Augmentation and Normalization.

Fig. 6 shows the data results of the data augmentation, it can be seen that the generated images change their position within the data augmentation process.

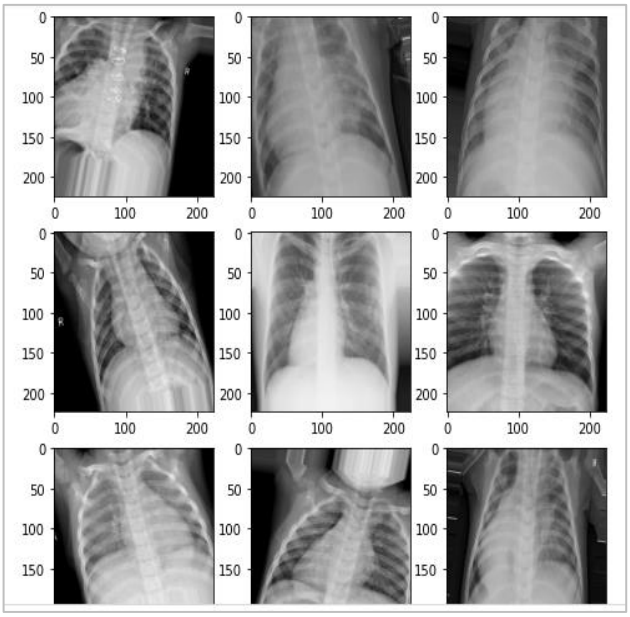


Fig. 6. Data Enhancement Process.

The use of the trained DenseNet neural network can be seen in Fig. 7.

The architecture of the new classifier using the pretrained DenseNet model is shown in Fig. 8.

Finally, the training of the convolutional Neural Network is shown in Fig. 9.

```
Model: "densenet201"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_2 (InputLayer)            (None, 224, 224, 3)  0
__________________________________________________________________________________________________
conv1/conv (Conv2D)             (None, 112, 112, 64) 9408        zero_padding2d_3[
__________________________________________________________________________________________________
conv1/bn (BatchNormalization)   (None, 112, 112, 64) 256         conv1/conv[0][0]
__________________________________________________________________________________________________
conv1/relu (Activation)         (None, 112, 112, 64) 0           conv1/bn[0][0]
__________________________________________________________________________________________________
zero_padding2d_4 (ZeroPadding2D (None, 114, 114, 64) 0           conv1/relu[0][0]
__________________________________________________________________________________________________
pool1 (MaxPooling2D)            (None, 56, 56, 64)   0           zero_padding2d_4[
```

Fig. 7. Download of the DenseNet Neural Network.

```
Arquitectura final:
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
densenet201 (Model)          (None, 7, 7, 1920)        18321984
_________________________________________________________________
global_average_pooling2d_3 ( (None, 1920)              0
_________________________________________________________________
dense_5 (Dense)              (None, 1000)              1921000
_________________________________________________________________
dropout_3 (Dropout)          (None, 1000)              0
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 1001
=================================================================
Total params: 20,243,985
Trainable params: 1,922,001
Non-trainable params: 18,321,984
_________________________________________________________________
```

Fig. 8. New Classifier Architecture Created with the Help of the Pre-trained DenseNet Model.

```
%%time

epochs=20

# Entrenar
history = model.fit_generator(training_set_imgs,
                              epochs=epochs,
                              steps_per_epoch=np.ceil(num_imgs_training/batch_size),
                              validation_data=testing_set_imgs,
                              validation_steps=np.ceil(num_imgs_testing/batch_size))

    Epoch 1/20
    9/9 [==============================] - 17s 2s/step - loss: 0.1268 - accuracy: 0.9460
    Epoch 2/20
    9/9 [==============================] - 7s 767ms/step - loss: 0.1130 - accuracy: 0.96
    Epoch 3/20
    9/9 [==============================] - 11s 1s/step - loss: 0.1152 - accuracy: 0.9432
    Epoch 4/20
```

Fig. 9. Artificial Neural Network Training.

b) Sprint N°2: In this phase, the Diagnostic Interface of the mobile application called "Diagnost" visualized in Fig. 10(c) was developed, the model created was saved in a disk, and finally, the Centos server was configured.

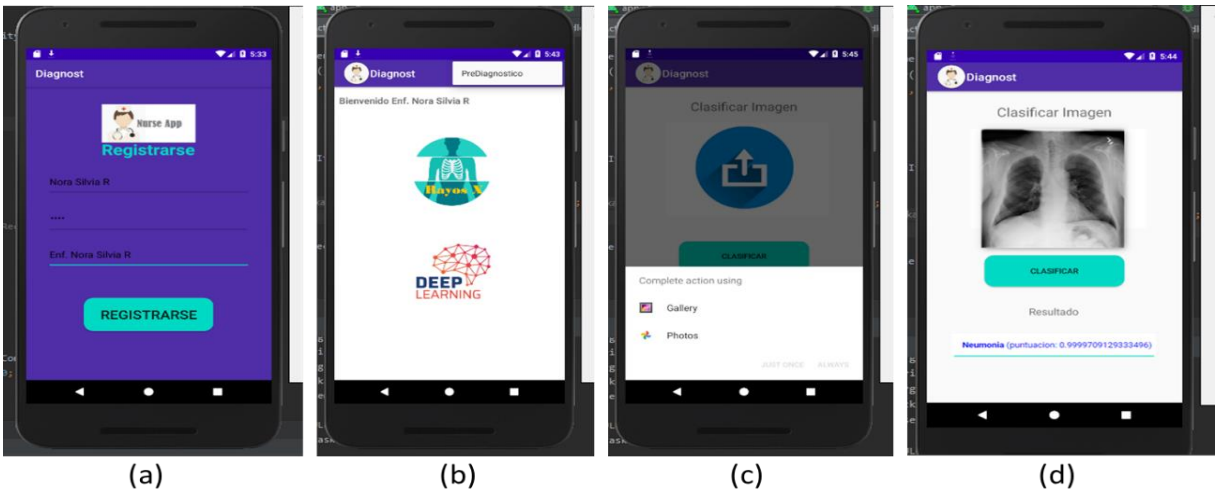


Fig. 10. Interfaces of the Developed Mobile Application "Diagnost".

Fig. 11 shows the server configuration, the model trained within the Deep environment was downloaded and the scripts to perform the requests were added.

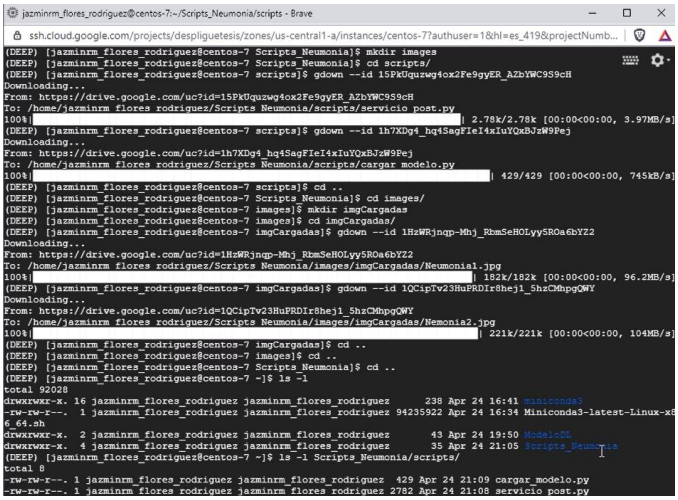


Fig. 11. Server Configuration.

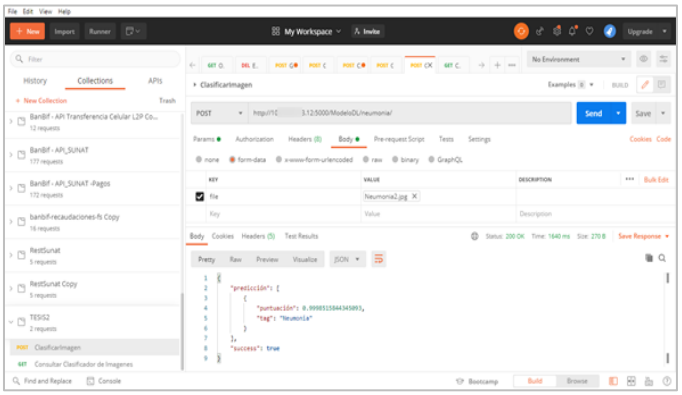


Fig. 12. Post Request from Postman.

Fig. 12 shows the validation of the request using the Postman Tool.

c) Sprint N°3: For this phase, the Login interface of the application was developed as shown in Fig. 10(a); the main menu of the application was created as shown in Fig. 10(b); and requests were made from the application as shown in Fig. 10(d).

VI. Results

A. Descriptive Analysis Experiment Group

Fig. 13 shows the comparison of means in post and pre groups of experiments for Indicator N° 1 Detection time:

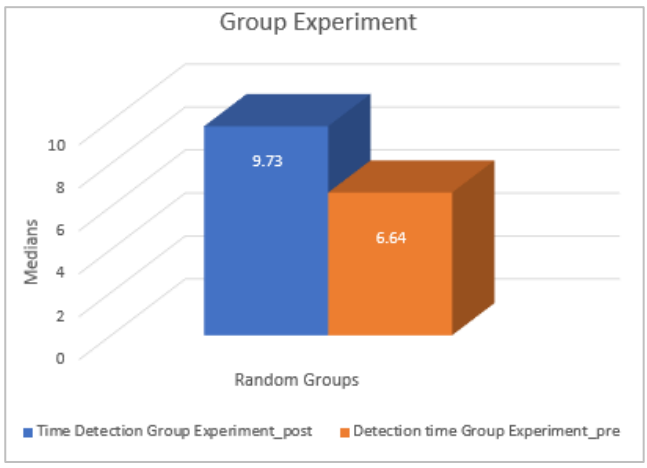


Fig. 13. Experiment Group Post and Pre Detection Time.

Fig. 14 shows the comparison of means in groups of post and pre-experiment groups for Indicator N° 2 Result.

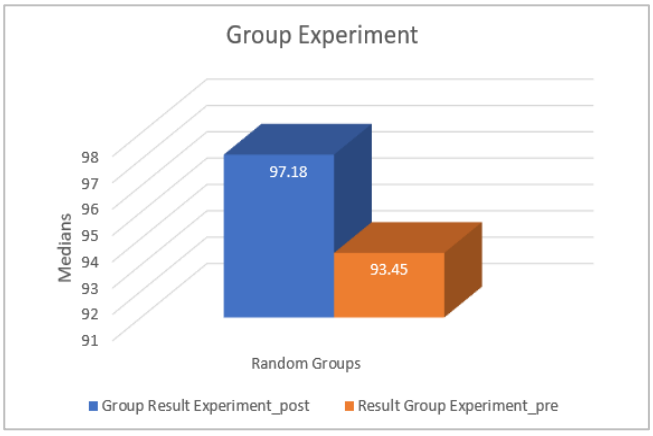


Fig. 14. Group Experiment Post and Pre Result.

Fig. 15 shows the comparison of Means in Groups of Experiment post and pre of Indicator N° 3 Attention Time:

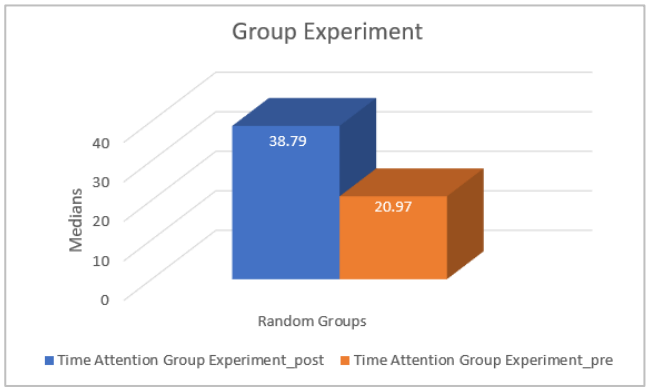


Fig. 15. Experiment Group Pre and Post-Time of Attention.

B. Inferential Results

1) Normality test: We proceeded to perform the normality test assisted by the SPSS software for the indicators of Detection Time, Result, and Attention Time through the Shapiro Wilks method, due to the fact that our sample size is less than 50.

Where:

Sig.<0.05 Adopts a non-normal distribution.

Sig.>= 0.05 Normal distribution.

Sig. P: Value or critical level of the contrast.

a) Indicador N°1: Para el Indicador se obtuvieron los siguientes resultados:

TABLE V. Normality Test Detection Time Indicator

	Kolmogorov-Smirnov			***Shapiro-Wilk***		
	Statistician	gl	Sig.	Statistician	gl	Sig.
Control Group	.299	33	<.001	.756	33	<.001
Experiment Group	.263	33	<.001	.872	33	.001

Table V indicates that the significance of the Control group is 0.001 and the significance of the Experiment group is 0.001, in both values are less than 0.05, then it is stated that the data have a normal distribution.

b) Indicator N°2: The following results were obtained for this indicator:

TABLE VI. Normality Test Indicator Result

	Kolmogorov-Smirnov			***Shapiro-Wilk***		
	Statistician	gl	Sig.	Statistician	gl	Sig.
Control Group	.329	33	<.001	.502	33	<.001
Group Experiment	.446	33	<.001	.404	33	<.001

Table VI indicates that the significance of the Control group is 0.001 and the significance of the Experiment group is 0.001, both values are less than 0.05, so the data are said to have a normal distribution.

c) Indicator N°3: The following results were obtained for this indicator:

TABLE VII. Normality Test Attention Time Indicator

	Kolmogorov-Smirnov			***Shapiro-Wilk***		
	Statistician	gl	Sig.	Statistician	gl	Sig.
Attention Time Control Group	.329	33	<.001	.502	33	<.001
Attention Time Experiment Group	.446	33	<.001	.404	33	<.001

The significance level for the control and experimental groups shown in Table VII is 0.001, both values are less than 0.05, so the data have a normal distribution.

2) Hypothesis Testing

a) Indicador N°1: H0, the Convolutional Neural Network-based mobile application does not allow early detection of Pneumonia in 2021. H1, The Convolutional Neural Network-based mobile application enables early detection of Pneumonia in 2021.

For the hypothesis test, the nonparametric Wilcoxon test was performed, where a significance level of 0.01 was obtained, which is less than 0.05, the limit value to see if the research is accepted.

In this case, by obtaining a p-value greater than 0.01, the alternative hypothesis (H1) is accepted and the null hypothesis (H0) is rejected.

b) Indicator N°. 2: H0, the Convolutional Neural Network-based mobile application does not allow accurate detection of Pneumonia in 2021. H1, the Convolutional Neural Network-based mobile application allows the accurate detection of Pneumonia in 2021.

For the hypothesis test, the nonparametric Wilcoxon test was performed, where a significance level of 0.340 was obtained, which is greater than 0.05, the limit value to see if the research is accepted.

In this case, by obtaining a p-value greater than 0.05, the null hypothesis (H0) is accepted and the alternative hypothesis (H1) is rejected.

c) Indicator N°3: H0, the mobile application based on the Convolutional Neural Network does not allow the reduction of medical assistance when diagnosing Pneumonia in 2021. H1, the mobile application based on the Convolutional Neural Network allows the reduction of medical assistance when diagnosing Pneumonia in 2021.

For the hypothesis test, the nonparametric Wilcoxon test was performed, where a significance level of 0.01 was obtained, which is less than 0.05, the limit value to see if the research is accepted.

In this case, by obtaining a p-value of 0.01, the alternative hypothesis (H1) is accepted and the null hypothesis (H0) is rejected.

VII. Discussion

From the results obtained in the present research work, it is observed in the descriptive analysis in the Detection Time indicator, Fig. 16 shows a deviation of 2.24 in the Control group and in the Experiment group (Fig. 17), a deviation of 1.78 is observed. Likewise, as observed in the hypothesis test, it has a significance level equal to 0.01, which is less than 0.05; determining that a mobile application based on the convolutional neural network allows the early detection of pneumonia in the year 2021. According to the research [30] for the diagnosis and treatment of pneumonia in pigs, it is verified that its expert system supports in some way reducing the time in which a pig is treated and that it is treated efficiently, in order to avoid aggravation and death, with the help of the knowledge provided by the specialist and thus develop an expert system that helps in making decisions to treat the disease in time.

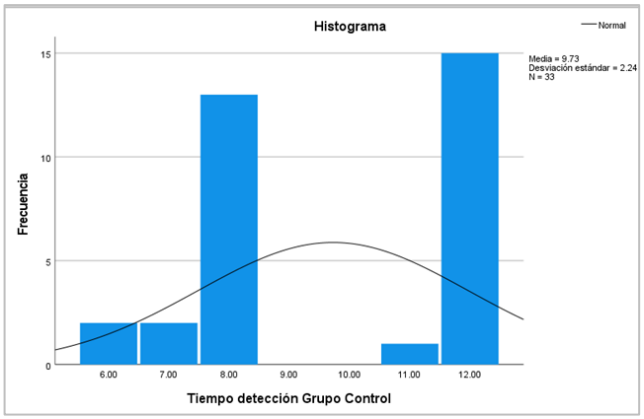


Fig. 16. Histogram of Control Group Indicator Time of Detection Indicator.

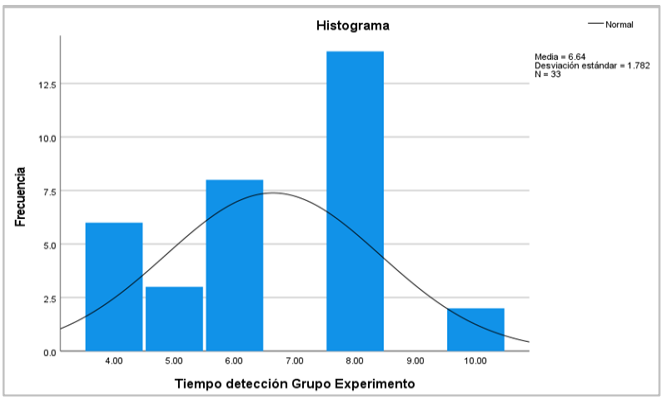


Fig. 17. Histogram of Experiment Group Indicator Detection Time.

Likewise, in the descriptive analysis in the Result indicator, in Fig. 18, a deviation of 4.10 is observed in the Control group, and in the Experiment group (Fig. 19), a deviation of 14.93 is observed. Likewise, as observed in the hypothesis test, s has a significance level equal to 0.340, which is greater than 0.05; determining that a mobile application based on the convolutional neural network does not allow the accurate detection of Pneumonia in the year 2021.

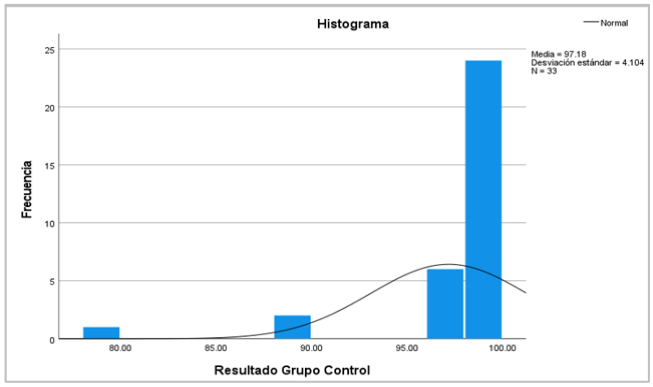


Fig. 18. Histogram of Control Group Indicator Result.

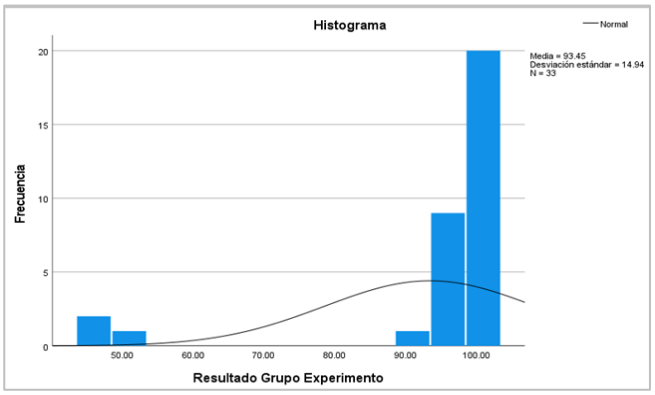


Fig. 19. Histogram of Experiment Group Indicator Result.

In addition, the descriptive analysis in the Attention Time indicator, in Fig. 20, shows a deviation of 12.38 in the Control group, and in the Experiment group (Fig. 21), a deviation of 9.16 is observed. Likewise, as observed in the hypothesis test s has a significance level equal to 0.01, which is less than 0.05; determining that a mobile application based on the convolutional neural network allows the reduction of medical assistance when making the diagnosis of Pneumonia, in the year 2021.

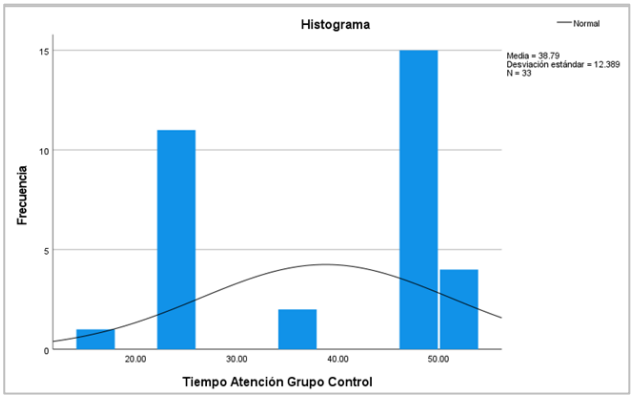


Fig. 20. FHistogram of Control Group Time to Care Indicator.

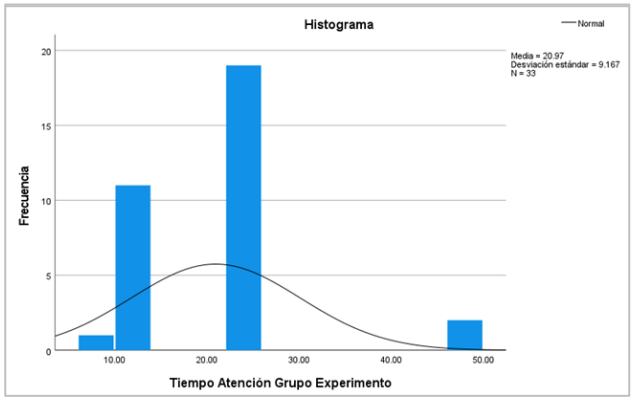


Fig. 21. Histogram of Experiment Group Indicator Time of Attention Indicator.

VIII. Conclusion

Convolutional network-based X-ray image classification has been widely used for different applications and predictions. In the present research work, we proposed the training of a Convolutional Neural Network for the Diagnosis of Pneumonia through the analysis of chest X-ray images implemented in a mobile application, in order to provide a technological tool to medical personnel, contributing to the early diagnosis of pneumonia.

From the results obtained, it was concluded that it was possible to prove that a mobile application based on the convolutional neural network allows the early detection of pneumonia and also allows the reduction of medical assistance when making the diagnosis. Although favorable results were achieved, it was also determined that there is still work to be done on the accuracy, since according to the studies for indicator 2, the mobile application based on the convolutional neural network does not allow the accurate detection of pneumonia, so for future research, it is proposed to improve the model to achieve greater accuracy.

References

[1] Forum of International Respiratory Societies, The Global Impact of Respiratory Disease, Third Edition. 2021.

[2] A. Kozinska, K. Wegrzynska, M. Komiazyk, J. Walory, I. Wasko, and A. Baraniak, "Viral Etiological Agent(s) of Respiratory Tract Infections in Symptomatic Individuals during the Second Wave of COVID-19 Pandemic: A Single Drive-Thru Mobile Collection Site Study," Pathogens, vol. 11, no. 4, p. 475, Apr. 2022, doi: 10.3390/pathogens11040475.

[3] World Health Organization (WHO), "Pneumonia," 2021. https://www.who.int/news-room/fact-sheets/detail/pneumonia (accessed Jul. 22, 2022).

[4] UNICEF, World Health Organization, World Bank Group, and United Nations Child, "Levels and Trends in Child Mortality," 2020.

[5] A. Balasubramanian, K. Ramalingam, A. Akash, E. Abinaya, and A. Abishek, "Review on Bacterial Pathogens Associated with Community-Acquired Pneumonia In Children," Pharmacologyonline, vol. 3, pp. 883–891, Dec. 2021.

[6] H.-C. Lin, C.-C. Lin, C.-S. Chen, and H.-C. Lin, "Seasonality of Pneumonia admissions and its association with climate: An eight-year nationwide population-based study," Chronobiol Int, vol. 26, no. 8, pp. 1647–1659, Dec. 2009, doi: 10.3109/07420520903520673.

[7] D. Lieberman, D. Lieberman, and A. Porath, "Seasonal variation in community-acquired pneumonia," Eur Respir J, vol. 9, no. 12, pp. 2630–2634, 1996, doi: 10.1183/09031936.96.09122630.

[8] J. Flores-Rodriguez and M. Cabanillas-Carbonell, "Mobile application for registration and diagnosis of respiratory diseases: A review of the scientific literature between 2010 and 2020," 2020 8th E-Health and Bioengineering Conference, EHB 2020, Oct. 2020, doi: 10.1109/EHB50910.2020.9280282.

[9] P. y C. de E. Centro Nacional de Epidemiologia, "Número de episodios de neumonías en menores de 5 años, Perú 2017 – 2022," MINSA , 2022. https://www.dge.gob.pe/portal/docs/vigilancia/sala/2022/SE08/neumonias.pdf (accessed Jul. 22, 2022).

[10] L. Andrade-Arenas and C. Sotomayor-Beltran, "Evolution of acute respiratory infections in Peru: A spatial study between 2011 and 2016," Proceedings of the 2019 IEEE 1st Sustainable Cities Latin America Conference, SCLA 2019, Aug. 2019, doi: 10.1109/SCLA.2019.8905563.

[11] J. Padilla, N. Espíritu, E. Rizo-Patrón, and M. C. Medina, "Neumonías en niños en el Perú: Tendencias epidemiológicas, intervenciones y avances," Revista Médica Clínica Las Condes, vol. 28, no. 1, pp. 97–103, Jan. 2017, doi: 10.1016/J.RMCLC.2017.01.007.

[12] R. del Pilar Nuñez-Delgado, R. Fredy Tapia-Pérez, E. Cachicatari-Vargas, R. Maritza Chirinos-Lazo, H. Daniel Alcides Carrión, and H. Carlos Alberto Seguin Escobedo, "Neumonía adquirida en la comunidad como factor de riesgo para enfermedades cardiovasculares," Revista del Cuerpo Médico Hospital Nacional Almanzor Aguinaga Asenjo, vol. 15, no. 1, pp. 35–41, Mar. 2022, doi: 10.35434/RCMHNAAA.2022.151.1072.

[13] G. Sun, T. Matsui, S. Kim, and O. Takei, "KAZEKAMO: An infection screening system remote monitoring of multiple vital-signs for prevention of pandemic diseases," 2014 IEEE 3rd Global Conference on Consumer Electronics, GCCE 2014, pp. 225–226, 2014, doi: 10.1109/GCCE.2014.7031086.

[14] G. Ricci et al., "Una aplicación móvil para pacientes con la enfermedad de Pompe y sus posibles aplicaciones clínicas," Neuromuscular Disorders, vol. 28, no. 6, pp. 471–475, 2018, doi: 10.1016/j.nmd.2018.03.005.

[15] R. Dharwadkar and N. A. Deshpande, "A Medical ChatBot," International Journal of Computer Trends and Technology, vol. 60, no. 1, pp. 41–45, 2018, doi: 10.14445/22312803/ijctt-v60p106.

[16] T. Karatekin et al., "Interpretable Machine Learning in Healthcare through Generalized Additive Model with Pairwise Interactions (GA2M): Predicting Severe Retinopathy of Prematurity," in 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), Aug. 2019, pp. 61–66. doi: 10.1109/Deep-ML.2019.00020.

[17] T. Karatekin et al., "Interpretable Machine Learning in Healthcare through Generalized Additive Model with Pairwise Interactions (GA2M): Predicting Severe Retinopathy of Prematurity," Proceedings - 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications, Deep-ML 2019, pp. 61–66, 2019, doi: 10.1109/Deep-ML.2019.00020.

[18] E. Kesim, Z. Dokur, and T. Olmez, "X-ray chest image classification by a small-sized convolutional neural network," 2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, EBBT 2019, pp. 1–5, 2019, doi: 10.1109/EBBT.2019.8742050.

[19] Mohammad S. Majdi1, K. N. Salman, M. F. Morris, N. C. Merchant, and Jeffrey J. Rodriguez, "Deep Learning Classification of Chest X-Ray Images," pp. 1–4, 2020.

[20] H. T. Nguyen, T. Bao, H. Hoang, T. Phuoc, and N. Cong, "Viral and Bacterial Pneumonia Diagnosis via Deep Learning Techniques and Model Explainability," International Journal of Advanced Computer Science and Applications, vol. 11, no. 7, p. 2020, 2020, doi: 10.14569/IJACSA.2020.0110780.

[21] T. Alaoui et al., "Classification of chest pneumonia from x-ray images using a new architecture based on ResNet," 2021.

[22] J. Durán Suárez and A. Del Real Torres, "Redes Neuronales Convolucionales en R Reconocimiento de caracteres escritos a mano Redes Neuronales Convolucionales en R Reconocimiento de caracteres escritos a mano Redes Neuronales Convolucionales en R," p. 78, 2017.

[23] A. Krizhevsky and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," pp. 1–9.

[24] J. L. Jameson, A. S. Fauci, D. L. Kasper, S. L. Hauser, D. L. Longo, and J. Loscalzo, Harrison. Principios de Medicina Interna , 20th ed. Access Medicina, 2020.

[25] R. Marroquin Peña, "Metodología de la Investigación," Universidad Nacional de Educación Enrique Guzmán y Valle, pp. 1–26, 2013.

[26] S. A. Alonso, S. L. García, R. I. León, G. G. Elisa, G. Á. Belén, and B. L. Ríos, "Métodos de investigación de enfoque experimental," Metodología de la investigación educativa, pp. 167–193, 2012.

[27] R. Hernandez, C. Fernandez, and P. Baptista, Metodologías de la Investigación. 2014.

[28] T. Dimes, "Conceptos Básicos De Scrum: Desarrollo De Software Agile Y Manejo De Proyectos Agile," p. 48, 2015.

[29] K. Schwaber and J. Sutherland, "La Guía de Scrum," 2013.

[30] A. S. Veloza Rodriguez, "Sistema experto de apoyo para el diagnóstico y tratamiento de la neumonía en cerdos," Scientia et technica, vol. 22, no. 1, p. 69, 2017, doi: 10.22517/23447214.12761.

Parameter Estimation in Computational Systems Biology Models: A Comparative Study of Initialization Methods in Global Optimization

Muhammad Akmal Remli[1]
Department of Data Science
Universiti Malaysia Kelantan, City Campus, Pengkalan Chepa, 16100
Kota Bharu, Kelantan, Malaysia

Nor-Syahidatul N.Ismail[2], Noor Azida Sahabudin[3]
Faculty of Computing, College of Computing and Applied Science
Universiti Malaysia Pahang
Pekan 26600, Pahang, Malaysia

Nor Bakiah Abd Warif [4]
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia
Parit Raja 86400, Johor, Malaysia

***Abstract*—This paper compares different initialization methods and investigates their performance and effects on estimating kinetic parameters' value in models of biological systems. Estimating parameters values is difficult and time-consuming process due to their highly nonlinear and huge number of kinetic parameters involved. Global optimization method based on an enhanced scatter search (ESS) algorithm is a suitable choice to address this issue. However, despite its resounding success, the performance of ESS may decrease in solving high dimension problem. In this work, several choices of initialization methods are compared and experimental results indicated that the algorithm is sensitive to the initial value of kinetic parameters. Statistical results revealed that uniformly distributed random number generator (RNG) and controlled randomization (CR) that being used in ESS may lead to poor algorithm performance. In addition, the different initialization methods also influenced model accuracy. Our proposed methodology shows that initialization based on opposition-based learning scheme have shown 10% better accuracy in term of cost function.**

***Keywords*—Metaheuristic; opposition-based learning; kinetic parameters; initialization method; metabolic engineering**

I. Introduction

Kinetic models of living cells have drawn the attention of both practitioners and researchers in recent years [1]. Their applications are important in metabolic and bioprocess engineering as they facilitate scholars to better understand, accurately predict and consistently improve the desired products in systems biology [2,3]. The models are formulated by means of ordinary differential equations (ODEs) to mimic various functional behaviours such as glycolysis reactions via metabolic pathway and phosphorylationin signal transduction of human cells. Due to the highly nonlinear biological systems, building such model is considered both challenges and time-consuming [4].

One important aspect of model building is parameter estimation, which consists of finding the best possible value of kinetic parameters that produce best fit model to the experimental data. The goodness of fit can be measured by minimizing distance value in the simulated model and experimental data. Thus, searching best parameter values in kinetic model can be depicted as a nonlinear optimization problem [5] and this class of problem is difficult to be solved. In this view, various optimization algorithms have been proposed in parameter estimation and their findings revealed that local optimization often fails to obtain snear-optimal solution [6]. Although improvements such as iterated local search have been proposed, they still consume high computational cost. Consequently, global optimization which is based on metaheuristic is an ideal option to address this issue. Global methods are quite capable in parameter estimation problem as they are more likely to reach the global minimum compared to local methods.

Enhanced scatter search (ESS) is one of the metaheuristic algorithm which have recently shown to yield promising outcomes in biological problems [7,8]. The algorithm benefits from global exploration and local exploitation using various choices of local search. The balanced tradeoff between global and local methods in ESS has shown promising results in solving optimization problems. However, when dealing with high dimension problem involving hundreds of kinetic parameters, performance of most global methods including ESS are deteriorate. One of the most neglected mechanisms in global methods is the way they generate the initial solution which were commonly derived using random number generator (RNG). The initialization methods may influence the efficiency and performance of the optimization algorithm in terms of its probability in finding the global minima, convergence's rate and variance of statistical results [9]. To date, only a few works have been done for comparing initialization methods in optimization. So far, no comparative study with regard to initialization method has been done in large-scale parameter estimation problem, particularly in the biological domain. The limitation of existing work in the field of global optimization is they only rely on RNG for initialization and only focus on search operator or the way new solution are produced. The high complexity of the problem such as in biological domain or healthcare is challenging and applying optimization method must properly select the best initialization because it will influence the

output. Hence, this issue motivates this research to further investigate the effect of different initialization method.

This paper compares and investigates the effects of several initialization methods (also known as diversification generation method in ESS algorithm) from the context of parameter estimation in systems biology models. The evolutionary algorithm based on ESS is utilized in this study due to its efficiency and reliability in parameter estimation problem [7]. The paper is organized as follows: Section II explicates the problem statement in parameter estimations; Section III delineates ESS algorithm; Section IV introduces several initialization methods; Section V compares the methods and presents the discussion of their results and Section VI presents the conclusion of this study.

II. Problem Background

In a nonlinear kinetic model of biological systems, the parameter estimation problem deals with finding an unknown value of kinetic parameters to minimize a distance (objective or cost function) between simulated model and real data. The value of cost function determines the goodness of fit of the model. The observables, which are referred to as the output state variable, are experimentally measured. The cost function of this problem, which is also known as weighted nonlinear least squares J is defined as:

$$J = \sum_{exp=1}^{n_{exp}} \sum_{obs=1}^{n_{exp}^{obs}} \sum_{s=1}^{n_s^{exp,obs}} (ym_s^{exp,obs} - y_s^{exp,obs}(\boldsymbol{p}))^T W (ym_s^{exp,obs} - y_s^{exp,obs}(\boldsymbol{p})) \quad (1)$$

where n_{exp} is the number of experiments, n_{exp}^{obs} is the number of observables per experiment and $n_s^{exp,obs}$ is the number of samples per observable in each experiment. Time series experimental data is denoted as $ym_s^{exp,obs}$ and predicted model is denoted as $y_s^{exp,obs}(\boldsymbol{p})$. The kinetic parameters vector to be estimated is $\boldsymbol{p}$. The time span for observables is denoted as T and finally W represents the weight matrix to balance the contributions of the observables. Minimization of the above cost function is subject to the following constraints:

$$\dot{x} = f(x, \boldsymbol{p}, t) \quad (2)$$

$$x(t_0) = x_0 \quad (3)$$

$$y = g(x, \boldsymbol{p}, t) \quad (4)$$

$$\boldsymbol{p}^{lb} \leq \boldsymbol{p} \leq \boldsymbol{p}^{ub} \quad (5)$$

where derivative of $\dot{x}$ is the function f of ODEs model that describes the dynamics of biological systems, x_0 is the initial condition at time t_0, g is the observable functions and $\boldsymbol{p}^{lb}$ and $\boldsymbol{p}^{ub}$ are the lower bound and upper bound of the kinetic parameter vector $\boldsymbol{p}$ respectively. This nonlinear and multimodal problem consists of many local minima. Thus, the process of finding the global minima is both challenging and time-consuming.

III. Enhanced Scatter Search (ESS) Algorithm

An enhanced scatter search (ESS) is a metaheuristic that belongs to the family of evolutionary algorithms. This algorithm is similar with genetic algorithm (GA) with regards to maintaining and updating their population members and evaluating their cost function in an iterative cycle. However, unlike GA, ESS does not use crossover and mutation as their evolutionary operators. Instead, it uses the combination among members in a reference set (*RefSet*). In this study, four phases of ESS algorithm are used, namely: 1) initialization method, 2) *RefSet* update method, 3) *RefSet* member generation and combination method, and 4) hybrid of the local search method. More advance designs and their mechanism can be found in [10,11].

This algorithm starts with randomly generating m population of diverse vectors by means of initialization (diversification generation) method. The m size is ten times the problem size to ensure that the large initial solutions in the search space are widely sampled, thus increasing the chances of avoiding local minima. Although uniformly distributed random number is a popular method usually utilized to generate initial solutions in various optimization algorithms, there are other strategies that may provide better initial solutions. Therefore, we compared and investigate four different initialization methods which will be briefly discussed in Section IV.

After the diverse vectors are generated, each vector is evaluated and half of the *RefSet* members $b/2$ (b is the *RefSet* size) is formed. The diversification method produces high quality initial *RefSet* member. The remaining *RefSet* members are choosen from the *RefSet* by random cycle to complete a *RefSet*. Then, the subset generation produces pairs of members in *RefSet*. Let us consider members of a *RefSet*, x^i, to be combined with the rest of members in *RefSet*, $x^j, \forall i, j \in [1,2,..b], i \neq j$. The pairs of the combination ($combi_1$ and $combi_2$) are defined as follows:

$$combi_1 = x^i - m(1 + \gamma \cdot \delta) \quad (6)$$

$$combi_2 = x^i + m(1 - \gamma \cdot \delta) \quad (7)$$

where

$$m = \frac{x^j - x^i}{2} \quad (8)$$

$$\gamma = \begin{cases} 1 \; if \; i < j \\ -1 \; if \; j < i \end{cases} \quad (9)$$

and

$$\delta = \frac{|j-i|-1}{b-2} \quad (10)$$

Every pair of combination ($combi_1$ and $combi_2$) in the *RefSet* members is used to create new hyper-rectangles which are defined by their relative positions and distance and thus, resulting in a new solution within them. The hyper-rectangles based combination methods are applied and are defined in the following equation:

$$x^{new} = combi_1 + R \cdot (combi_2 - combi_1) \quad (11)$$

where x^{new} is new solution generated and R is the random number, $R \sim U([0,1])$. This combination strategy is similar to the mutation operator in differential evolution (DE) [12,13], which is effective in updating population members. In ESS,

the vectors of combination ($combi_1$ and $combi_2$) are systematically generated. They are not randomly generated, as practiced in DE. Using this combination strategy, every *RefSet* member generates a hyper-rectangle among the rest of *RefSet* member. The new number of solution produces $b-1$ solution for each *RefSet* member. After this strategy is implemented, a new solution (offspring) is generated with different distance and direction around their *RefSet* members (parents). In this case, if the offsprings have better (lesser) fitness value compared to their parents, the current solutions will be replaced. Otherwise, the same *RefSet* members will be used for the next iteration. In order to accelerate convergence, gradient local search is performed using Sequential Quadratic Programming (SQP). The algorithm is applied using *fmincon* solver in MATLAB. This solver minimizes the cost function using the results obtained in ESS using different vectors. If the solution obtained by *fmincon* outperforms the solution generated by ESS, the solution from *fmincon* will replace the current solution and it will in turn be added to *RefSet* members for further update. Otherwise, the solution from *fmincon* will be discarded. This process is repeated until the stopping criteria are met.

IV. Initialization Methods

We implement five initialization methods in the ESS algorithm in order to compare and investigate their effects on parameter estimation. The methods are random number generator (RNG), controlled randomization (CR), opposition-based learning (OBL), quasi-opposition learning (QOBL) and chaotic (Tent) map.

A. Random Number Generator (RNG)

The most commonly used initialization method in optimization algorithms is random number generator (RNG). RNG is defined as below: Let $X_i(x_{i,1}, x_{i,2}, \dots, x_{i,D})$ be the ith member of the population, each $x_{i,j}$ is generated between lower and upper bound (lb_i,ub_j). In summary, it generates uniformly distributed random numbers as in the following equation:

$$x_{i,j} = ub_j + R \cdot (lb_j - ub_j), j = 1, \dots, D \quad (12)$$

where R is the random numbers between 0 and 1. The vector of $X_{i,j}$ contains a list of random initial population generated between lower bound and upper bound $[lb, ub]$ for each of the variable.

B. Controlled Randomization (CR)

Unlike RNG, controlled randomization (CR) strategy generates the first five populations ($n = 5$) of equal size for each vector as in the following equation [14]:

$$x_{i,j} = \frac{R \cdot (npar) + i - 1}{n} \quad (13)$$

where $x_{i,j}$ is the vector of candidate solutions, R is the random numbers in the between 0 and1, and $npar$ is the number of kinetic parameters. After the first five vectors are generated, the remaining vectors are generated randomly and all initial solutions are put in the boundaries:

$$x_{new} = x_i \cdot (ub - lb) + lb \quad (14)$$

where lb and ub are lower and upper bounds, respectively. This strategy generates a set of diverse vectors which contain equal sizes of range in the first five vectors and other random vectors lie in sixth vector to m diverse vectors. It should be noted that ESS algorithm used CR strategy as its default initialization method [11].

C. Opposition-based Learning (OBL)

Opposition-based learning (OBL) is introduced in the field of computational intelligence [15]. This scheme is subsequently applied in optimization areas [16]. The basic idea of OBL is to generate a set of opposite numbers from first initial solutions generated by RNG, as follows: Let $x \in [lb, ub]$ is a random value. The opposition value of x is defined by:

$$\breve{x} = lb + ub - x \quad (15)$$

Based on Eq. (15), the opposite point for optimization in dimension space *D* is defined as follows:

Let $X_i(x_{i,1}, x_{i,2}, \dots, x_{i,D})$ be the i th member of the population and each member $x_{i,j}$ be bounded by (lb_i,ub_j) and $x_i \in [lb_i, ub_i], \forall i \in \{1,2,..D\}$. Thus, the opposite value of $\breve{X}_\iota = (\breve{x}_{i,1}, \breve{x}_{i,2}, .., \breve{x}_D)$ is defined as:

$$\breve{x}_{i,j} = lb_i +, ub_i - x_{i,j}, j = 1, \dots, D \quad (16)$$

Both X and $\breve{X}$ is merged. Now, let us assume $F(x)$ is the cost function in minimization problem. If cost function value of $f(\breve{x})$ is smaller than $f(x)$, $f(\breve{x}) < f(x)$, point X can be replaced with $\breve{X}$. Otherwise, point X will stay in the current population. All the population members will be evaluated and the initial population with fittest members among X and $\breve{X}$ is formed.

D. Quasi Opposition-based Learning (QOBL)

Another family of OBL is quasi-opposition learning (QOBL) which modified version of OBL that increases population uniformity [17]. Considering the opposite point in equation 16, the middle point $m_j = \{m_1, m_2, .., m_D\}$ is calculated as follows:

$$m_j = \frac{lb_j + ub_j}{2}, \forall j \in \{1,2,..D\}. \quad (17)$$

Then, quasi-opposite point $\breve{X}^Q = (x_1^Q, x_2^Q, .. x_n^Q)$ is selected randomly within the range of opposite points of $\breve{X}$ and middle point m:

$$\breve{x}_i^Q = \begin{cases} R \cdot (mj, \breve{x}_{i,j}) \; if \, x_{i,j} \le m_j \\ R \cdot (\breve{x}_{i,j}, mj) \; if \, x_{i,j} > m_j \end{cases} \quad (18)$$

where R is a random value drawn uniformly in the range of lower bound and upper bound. Like OBL, $\breve{X}^Q$ and X are merged and the best solution is chosen, that is, the fitness of the cost function.

E. Chaotic Map

Another alternative of uniformly distributed random numbers for diversification generation is chaotic map. This approach is based on the deterministic and chaotic systems and it is not necessarily random. In this paper, we investigate

one family of the chaotic map, which is Tent map [18] which is defined as:

$$x_{i,j}^{(k+1)} = \begin{cases} \frac{x_{i,j}^{(k)}}{0.7} x_{i,j} < 0.7 \\ \frac{10}{3}\left(1 - x_{i,j}^{(k)}\right) x_{i,j} \geq 0.7 \end{cases} \quad (19)$$

where $x_{i,j}^{(k+1)}$ is j th variable of i th individual in k th iteration. $x_{i,j}^{(k)}$ is the initial variable that is generated randomly using RNG. In this strategy, solutions which are generated from Tent map are not predictable and are highly sensitive to initial variables.

V. Result and Discussion

A large-scale model is used to test the different initialization methods in ESS algorithm. The model involves dynamic processes that reproduce the response to a pulse in extracellular glucose concentrations of central carbon metabolism (CCM) in *E. coli*. This model consists of 18 metabolites: 17 internal metabolites in cytosol and 1 extracellular metabolite (*glucose*) in extracellular compartment.

These metabolites consists of PEP, G6P, PYR, F6P, G1P, 6PG, FDP, GAP, CPEP, CG6P, CPYR, CF6P,GLCex, CG1P, CPG, CFDP, CGAP and Glucose. The model also contains 48 reactions coupled with 166 kinetic parameters. The mathematical formulation and description of this model can be found in [19]. Table I summarizes the characteristics of CCM E. coli model.

TABLE I. Characteristics of the Central Carbon Metabolism (CCM) in E.coli

Number of kinetic parameters	Dynamic metabolites	Observed metabolites	Noise level	Lower value	Upper value
116	18	9	Real	$0.1 \times p_{ori}$	$10 \times p_{ori}$

Note: For fair comparison, lower and upper bound are set as a function of p_{ori}, where p_{ori} is a set of kinetic parameters obtained from original publication. In this data, only observed metabolities are measured.

In order to obtain statistically significant result, we ran each initialization method discussed in Section IV, 20 times and reported the best, mean, and worst results; as well as average function evaluations, CPU time and standard deviation. Function evaluation for each run was limited to 100,000 (the stopping criteria) to let the algorithm obtain the best parameter values. The RefSet size used was 36, which is the recommended size in this problem. With the high number of function evaluations and hundreds of parameters, the minimization process is expected to consume very lengthy CPU time. To surmount this drawback, Parallel Computing Toolbox in MATLAB has been used and it expedited the computation by assigning each run to eight different processors (logical cores) simultaneously. In this strategy, eight computations for each method was run independently using parfor loop which is available from the abovementioned toolbox. It should be noted that a single run takes approximately 11 hours, so 20 runs take approximately 220 hour. Using the parallel strategy, 20 runs only take approximately 33 hour, which reduced 72.6% of CPU time needed. All methods were experimented on i7 CPU with 16GB RAM which implemented in MATLAB 2015.

Table II shows that the best (minimum) cost function was obtained from QOBL method with $J = 210.0511$. The second best value is 229.1855, which was obtained from CR. Only RNG, OBL and TENT produced cost function values which were slightly higher than the published benchmark value, 233.90. The results revealed that QOBL is the best method in finding global minimum. However, although QOBL presented the minimum value, its average standard deviation was relatively higher than RNG, OBL and TENT. RNG is the most consistent method followed by OBL, having 4.5955 and 4.7331 standard deviation each, respectively. In terms of search effort, RNG produced the lowest average of function evaluations (1.2327e+05) and also its CPU time is also the lowest with 3.8191e+04 seconds. It should be observed that QOBL is the best initialization method if we consider its ability in minimizing the cost function in large-scale parameter estimation problem.

TABLE II. Experimental Results Obtained from the 20 Runs Conducted using different Initialization Methods

Initialization method	Best value	Worst value	Mean value	Standard deviation	Function evaluation	CPU time (s)
RNG	234.6651	252.0459	245.7053	4.5955	1.2327e+05	3.8191e+04
CR (rerun)	229.1855	270.0339	245.5427	10.0586	1.2421e+05	4.3142e+04
OBL	234.5223	250.9120	243.9961	4.7331	1.2546e+05	4.1942e+04
QOBL	210.0511	255.0070	241.4211	9.3596	1.2830e+05	4.2574e+04
TENT	234.2876	259.8841	246.2248	6.4994	1.2533e+05	4.0755e+04

Note: The best (minimum) value of cost function (weighted nonlinear least square) is shown in shaded cell. CR (rerun) indicates our own experimental result (in case when comparing with publish result in the next subsection).

Additional information to compare the different initialization methods is given in Fig. 1 and Fig. 2. The figures depicts the best curves (with minimum cost function *J*) among the 20 runs obtained from RNG, CR, OBL, QOBL and Tent methods. The curves show that all methods are able to minimize the cost function at a similar rate in terms of function evaluations and CPU time. Note that we have set the same initial guess as the initial value for all methods. This gives fairer comparison and assumes that the search space is feasible. In Fig. 1, starting from the first fractions of 1,000 function evaluations, QOBL has better speed and found acceptable cost function value when it reached approximately 25,000 function evaluations. Meanwhile, CR has the slowest speed until it reached around 40,000 evaluations. All methods continued to progress they reached the final evaluations. In this case, QOBL found the best value of 210.051 at 123,979 function evaluations. At the end of the evaluations, all solutions were able to achieve equivalent solutions in terms of quality, although, the best (minimum) value was obtained by QOBL while the worst (maximum) value obtained by RNG. In

Fig. 2, default initialization method based on CR has obtained slow convergene rate compared to others. It can be noticed where CR obtained acceptable value of cost function when CPU time reached nearly 8 hours, while QOBL reached the acceptable value at nearly 1.3 hours. The results revealed that RNG which is mostly used initialization method and CR as default initialization method in ESS have obtained poor result. Meanwhile, initialization method based on QOBL is a better alternative which was not only able to speed up convergence, but also obtain optimal solution.

To evaluate the quality of the parameter estimates, we compared the best result (QOBL) with a published benchmark result [20]. One thing to note is benchmark result used conventional CR as their initialization methods for the ESS. Table III shows that our study produced the best cost function, J = 210.05 compared to the benchmark's J = 233.90. However, our work obtained a bigger number of function evaluations compared to CR, with a difference of 33,251 evaluations. In terms of CPU time, the benchmark also produced shorter time of approximately 3 hours for a single run. Due to different stopping criteria used (the benchmark study uses CPU time) and different hardware specifications, comparing QOBL and CR in terms of CPU time seems unfair. It should be noted that in terms of efficiency in finding global minimum, method used in this study produced better results compared to CR.

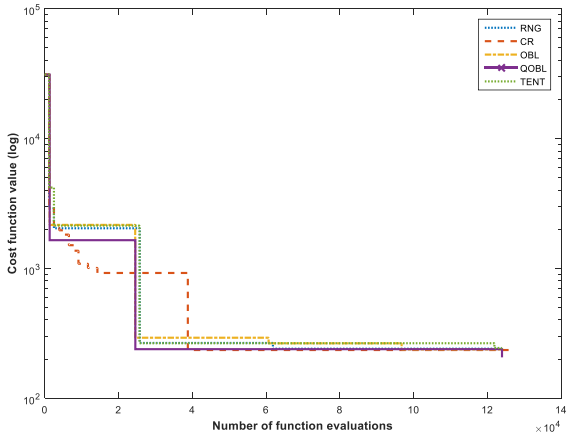


Fig. 1. Convergence of the Five Initialization Methods in Scatter Search.

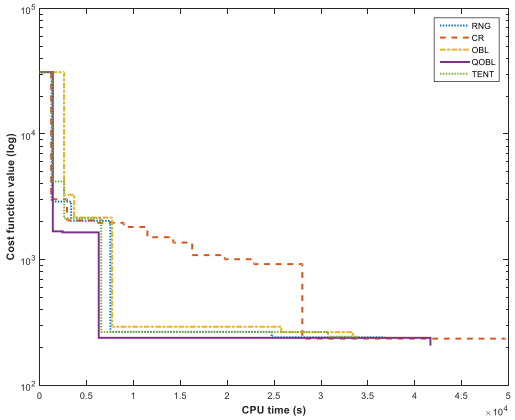


Fig. 2. Slow Convergence of the Five Initialization Methods in Scatter Search.

TABLE III. COMPARISON OF QOBL METHOD WITH PUBLISHED BENCHMARK IN SCATTER SEARCH

Initialization Methods	Best cost function J	Number of function evaluations	CPU Time (seconds)	ΣNRMSE
QOBL	210.05	12.3979e+04	4.1671e+04	2.3773
CR [20]	233.90	9.0728e+04	1.0800e+04	2.4921

Note: The best values are shown in shaded cell.

The decision variables (kinetic parameters) obtained from QOBL may provide the optimal model prediction since it produced the lowest cost function J and may produce best fit to experimental data. The goodness of fit can be measured by calculating root-mean-square-error ($RMSE$) for all metabolites. $RMSE$ is used to measure prediction error which is the different between experimental data and predicted model. The following equation defines $RMSE$,

$$RMSE = \sqrt{\frac{\sum_{exp=1}^{n_{exp}} \sum_{s=1}^{n_s^{exp,obs}} (ym_s^{exp,obs} - y_s^{exp,obs}(\boldsymbol{p}))^2}{n_{\epsilon xp} \cdot n_s^{exp,obs}}} \quad (20)$$

with the same notation defined in Eq. (1). In this case, normalized RMSE is used to cater for different magnitudes of observables. Each RMSE is divided by the range of value of observables as defined as,

$$NRMSE = \frac{RMSE}{\max(ym^{exp,obs}) - \min(ym^{exp,obs})} \quad (21)$$

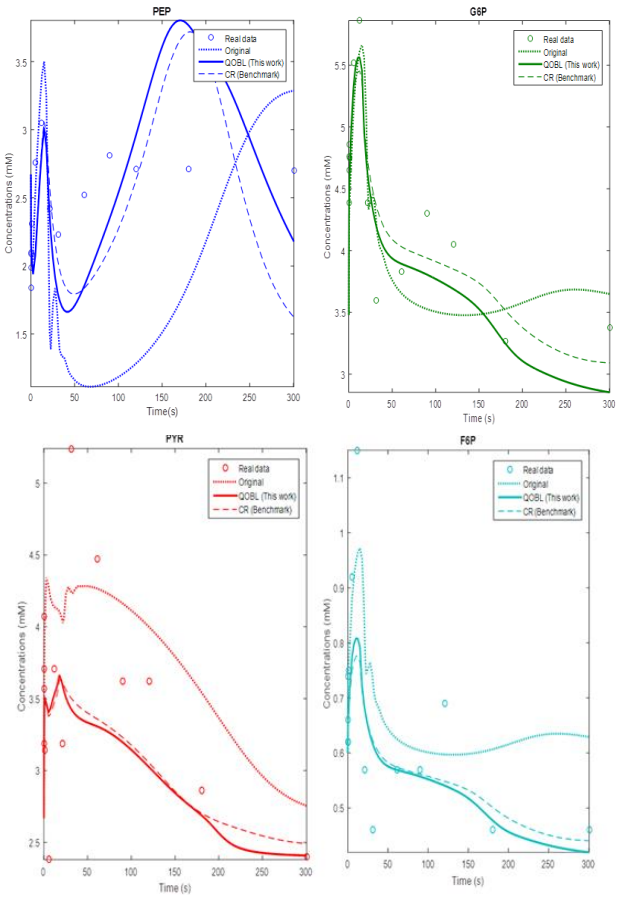


Fig. 3. Model Prediction over Experimental Data.

Thus, $\Sigma NRMSE$ is the sum of all $NRMSE$ for all observables. Table II shows our $\Sigma NRMSE$ is lower than the benchmark, means that it has better fit compared to the parameters obtained in the benchmark.

To measure the goodness of fit from parameter estimates with QOBL, we plot the model prediction over experimental data as shown in Fig. 3. For the sake of brevity, we plot only four out of nine metabolites. We chose metabolites which have very high nonlinear biological system, namely, PEP, G6P, PYR and F6P. The figure shows dynamic concentration change of extracellular glucose that responded to a pulse in central carbon metabolism. To compare the goodness of fit with other parameters, we also plot another fit based on parameters value from original published results and benchmark results. It should be observed that kinetic parameters retrieved from initialization methods based on QOBL represent a good fit between experimental data and predicted model.

VI. Conclusion

This paper studies different initialization methods and investigated their performance and effects on solving large-scale parameter estimation problem. We compared five initialization methods which are based on stochastic and randomization methods and implement them in ESS algorithm. Experimental results revealed that the choice of initialization (diversification generation) methods influenced the performance of the algorithm. The quality of solution, speed of convergence and statistical results were obtained with different characteristics derived from different initialization methods. Our statistical analyses revealed that the most popular initialization method, random number generator (RNG) performs poorly and there are significant better alternatives to this method, which have comparable computational requirements. In addition, the accuracy of model prediction also depends on the choice of initialization methods. Further investigation is needed to discover whether the same findings can be produced when different models and problems associated with parameter estimation are used. More intensive studies also need to be conducted on why some methods are generated more consistent performance in terms of statistical analysis and value of kinetic parameters in biological systems.

Acknowledgment

The authors would like to thank the Malaysian Ministry of Higher Education via the Fundamental Research Grant Scheme (FRGS), RACER/1/2019/ICT02/UMP//1 (University Reference RDU192601).

References

[1] Smallbone K, Mendes P. Large-Scale Metabolic Models: From Reconstruction to Differential Equations. Ind Biotechnol 2013;9:179–84. doi:10.1089/ind.2013.0003.

[2] Almquist J, Cvijovic M, Hatzimanikatis V, Nielsen J, Jirstrand M. Kinetic models in industrial biotechnology - Improving cell factory performance. Metab Eng 2014;24:38–60. doi:10.1016/j.ymben.2014.03.007.

[3] Hussain F, Jha SK, Jha S, Langmead CJ. Parameter discovery in stochastic biological models using simulated annealing and statistical model checking. Int J Bioinform Res Appl 2014;10:519–39. doi:10.1504/IJBRA.2014.062998.

[4] Link H, Christodoulou D, Sauer U. Advancing metabolic models with kinetic information. Curr Opin Biotechnol 2014;29:8–14. doi:10.1016/j.copbio.2014.01.015.

[5] Mendes P, Kell D. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics 1998;14:869–83. doi:10.1093/bioinformatics/14.10.869.

[6] Moles CG, Mendes P, Banga JR. Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Res 2003;13:2467–74. doi:10.1101/gr.1262503.

[7] Egea J a., Martí R, Banga JR. An evolutionary method for complex-process optimization. Comput Oper Res 2010;37:315–24. doi:10.1016/j.cor.2009.05.003.

[8] Mansour N, Kehyayan C, Khachfe H. Scatter search algorithm for protein structure prediction. Int J Bioinform Res Appl 2009;5:501–15. doi:10.1504/IJBRA.2009.028679.

[9] Kazimipour B, Li X, Qin AK. Initialization methods for large scale global optimization. 2013 IEEE Congr. Evol. Comput., 2013, p. 2750–7. doi:10.1109/CEC.2013.6557902.

[10] Egea J, Balsa-Canto E. Dynamic optimization of nonlinear processes with an enhanced scatter search method. Ind Eng Chem Res 2009;48:4388–401.

[11] Egea JA, Henriques D, Cokelaer T, Villaverde AF, MacNamara A, Danciu D-P, et al. MEIGO: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. BMC Bioinformatics 2014;15:136. doi:10.1186/1471-2105-15-136.

[12] Storn R, Price K. Differential Evolution -- A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. J Glob Optim 1997;11:341–59. doi:10.1023/A:1008202821328.

[13] Chong C, Mohamad M, Deris S, Shamsir M, Chai L, Choon Y. Parameter Estimation by Using an Improved Bee Memory Differential Evolution Algorithm (IBMDE) to Simulate Biochemical Pathways. Curr Bioinform 2014;9:65–75. doi:10.2174/15748936113080990007.

[14] Rodriguez-Fernandez M, Egea JA, Banga JR. Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. BMC Bioinformatics 2006;7:483. doi:10.1186/1471-2105-7-483.

[15] Tizhoosh HR. Opposition-Based Learning: A New Scheme for Machine Intelligence. Comput Intell Model Control Autom 2005 Int Conf Intell Agents, Web Technol Internet Commer Int Conf 2005;1:695–701. doi:10.1109/CIMCA.2005.1631345.

[16] Rahnamayan S, Tizhoosh HR, Salama MM. Opposition-based differential evolution. Stud Comput Intell 2008;143:155–71. doi:10.1007/978-3-540-68830-3_6.

[17] Rahnamayan S, Tizhoosh HR, Salama MMA. Quasi-oppositional differential evolution. 2007 IEEE Congr. Evol. Comput. CEC 2007, 2007, p. 2229–36. doi:10.1109/CEC.2007.4424748.

[18] Saremi S, Mirjalili S, Lewis A. Biogeography-based optimisation with chaos. Neural Comput Appl 2014;25:1077–97. doi:10.1007/s00521-014-1597-x.

[19] Chassagnole C, Noisommit-Rizzi N, Schmid JW, Mauch K, Reuss M. Dynamic modeling of the central carbon metabolism of Escherichia coli. Biotechnol Bioeng 2002;79:53–73. doi:10.1002/bit.10288.

[20] Villaverde AF, Henriques D, Smallbone K, Bongard S, Schmid J, Cicin-Sain D, et al. BioPreDyn-bench: a suite of benchmark problems for dynamic modelling in systems biology. BMC Syst Biol 2015;9:1–15. doi:10.1186/s12918-015-0144-4.

Determinants of Information Security Awareness and Behaviour Strategies in Public Sector Organizations among Employees

Al-Shanfari I[1], Warusia Yassin[2]*, Nasser Tabook[3], Roesnita Ismail[4], Anuar Ismail[5]
Department of Computer System & Communication[1, 2]
Universiti Teknikal Malaysia Melaka, Durian Tunggal, Melaka, MALAYSIA[1, 2]
College of Arts and Applied Science, Computer Science Department, Dhofar University, Salalah, OMAN[3]
Faculty of Science and Technology, Universiti Sains Islam Malaysia, Negeri Sembilan, MALAYSIA[4]
Ask-Pentest Sdn Bhd, Kuala Lumpur, MALAYSIA[5]

***Abstract*—In this digital era, protecting an organisation's sensitive information system assets against cyberattacks is challenging. Globally, organisations spend heavily on information security (InfoSec) technological countermeasures. Public and private sectors often fail to secure their information assets because they depend primarily on technical solutions. Human components create the bulk of cybersecurity incidents directly or indirectly, causing many organisational information security breaches. Employees' information security awareness (ISA) is crucial to preventing poor information security behaviours. Until recently, there was little combined information on how to improve ISA and how investigated factors influencing employees' ISA levels were. This paper proposed a comprehensive theoretical model based on the Protection Motivation Theory, the Theory of Planned Behaviour, the General Deterrence Theory, and Facilitating Conditions for assessing public sector employees' ISA intentions for information security behaviour. Using a survey and the structural equation modelling (SEM) method, this research reveals that the utilised factors are positively associated with actual information security behaviour adoption, except for perceived sanction certainty. The findings suggest that the three theories and facilitating conditions provide the most influential theoretical framework for explaining public sector employees' information security adoption behaviour. These findings support previous empirical research on why employees' information on security behaviours vary. Consistent with earlier research, these psychological factors are just as critical as facilitating conditions in ensuring more significant behavioural intention to engage in ISA activities, ensuring information security behaviour. The study recommends that public-sector organisations invest in their employees' applied information security training.**

Keywords*—Information security awareness; behaviour strategies; self-administered questionnaire; structural equation modelling (SEM)

I. Introduction

Securing information system assets has become a primary issue for organisations in today's digital environment to protect them from criminal assaults. In recent years, both cybercrime and data breaches have expanded considerably. By 2021, cyber-crime is predicted to cost more than $6 trillion, up from $3 trillion in 2015, according to the Cybersecurity Business Report [1]. As a result, organisations are constantly struggling to protect the security of their information assets, which causes them to spend heavily on technical countermeasures [2]. However, concentrating just on the technological areas of information security is insufficient since information security is multidisciplinary, with the human factor playing a significant role. The exploitation of human factors is responsible for a considerable percentage of organisational information security incidents [1]. In other respects, human error is directly or indirectly primarily the result of security breaches, including both intentional and unintentional negative behaviour [3]. According to ENISA [4], about 77% of data breaches occur due to human vulnerability. Additionally, it has been previously shown that over half of all information security breaches are caused by staff's insufficient compliance with information security policies [5].

In consideration of this context, staff members' information security awareness (ISA) has a significant influence on their information security behaviours and their compliance with security policies [6], [7]. Previous research has asserted that a lack of staff ISA as defined by information security policies (ISP) and procedures is the main reason for sensitive information misbehaviour [3]. Additionally, ISA has been a critical concern in research and practice [8] because humans are often identified as a weak link in efforts to protect systems and networks [9]. For this reason, among others, the most recent Cyber Security Breaches Survey 2019 demonstrates that cyber security is a top priority for senior management in the workplace [10].

Even though research and practice prioritise employees' information security awareness, most employees are unaware of information security risks and challenges [6]. For instance, about 90% of cybersecurity experts reported that the organisations for which they work feel exposed to insider threats [3]. According to Jaeger [11], research on ISA is still in its infancy, with numerous new areas to be explored. Even though many studies have been done on ISA, there is still no complete picture of the concept of ISA and how it fits into other constructs [11]. Other studies support this, suggesting that ISA campaigns and education fail to influence employees' behaviour for various reasons [12], [13].

*Corresponding Author.

It has been revealed that organisations fail with their ISA campaigns because they do not appropriately employ the factors impacting personnel's ISA levels while producing the content and developing material for the ISA campaigns [13], [1]. Most importantly, it was found that there were no good ways to make exciting and valuable materials for improving ISA. As a result, several behavioural factors, such as communication channels [14, 15], were not considered when ISA campaigns or initiatives were made to keep improving ISA levels [13].

Our assessment [67] of the relevant literature revealed that most research that relied on constructing models for ISA focused only on behavioural intentions or actual behaviour. Therefore, concentrating on both aspects is crucial and needs additional research [47], [51]. In ISA-related research, facilitating conditions factors have been mostly neglected; this issue also needs thorough investigation. This research implemented its developed model by concentrating on behavioural intention and actual behaviour and two facilitating conditions: organisational support and communication channels to fill these gaps. Incorporating these factors and verifying that they can enhance ISA by employing a combination of control, motivation, prediction, deterrence, and technical-related factors—which aid in managing human thought from a broad perspective to achieve optimal behavioural security practices—will enhance the current understanding.

This study, however, is a continuation of our prior research [16], which seeks to improve ISA among public-sector organisation staff by merging motivational, control/prediction, and deterrence variables into employees' behaviour to promote security awareness and reduce breaches. This study looks at the development and evaluation of a conceptual framework based on factors from the literature on information security from previous international studies. According to the model's constructs, the mediator variable is ISA's behavioural intention, and the dependent variable is InfoSec's actual behaviour. In contrast, the independent variables are a set of ten variables that have never been investigated together in the InfoSec literature. The theoretical background and conceptual model are described in Section II, followed by the methodology and results in Sections III and IV, respectively. Section V discusses the comparative evaluation of the study model. Finally, in Section VI, the conclusion is provided, along with limitations and suggestions for future studies.

II. Theoretical Background and Conceptual Model

This study highlights a new perspective relying on protection motivation theory (PMT), theory of planned behaviour (TPB), and general deterrence theory (GDT), as well as the facilitating conditions to enhance employees' ISA intentions. The different perspectives of these theories and the facilitating conditions show the whole chain of the InfoSec behaviour adoption process. Thus, it helps organisations reduce information security breaches by changing employees' behaviour to match information security policies and rules [7, 15, 17–22]. An assessment of theories utilised in related work revealed that the theories of TPB, PMT, and GDT are most often used [23]; [24]. TBP is one of the most influential theories describing human behaviour in different fields, such as organisational behaviour, public relations, healthcare, or advertising [11]. PMT is one of the most effective models for predicting a person's motivation and intention to take preventative measures [25]. GDT provides a practical focal point for describing misbehaviour [15]. The security education, training, and awareness (SETA) initiatives are the methods through which organisations raise information security awareness, educate staff on the necessity of ISA, and train end-users to take on information security activities [26]. Furthermore, facilitating conditions help employees accomplish their duties and responsibilities more quickly and effortlessly [22]. Fig. 1 presents the study model and utilised factors in a concise form.

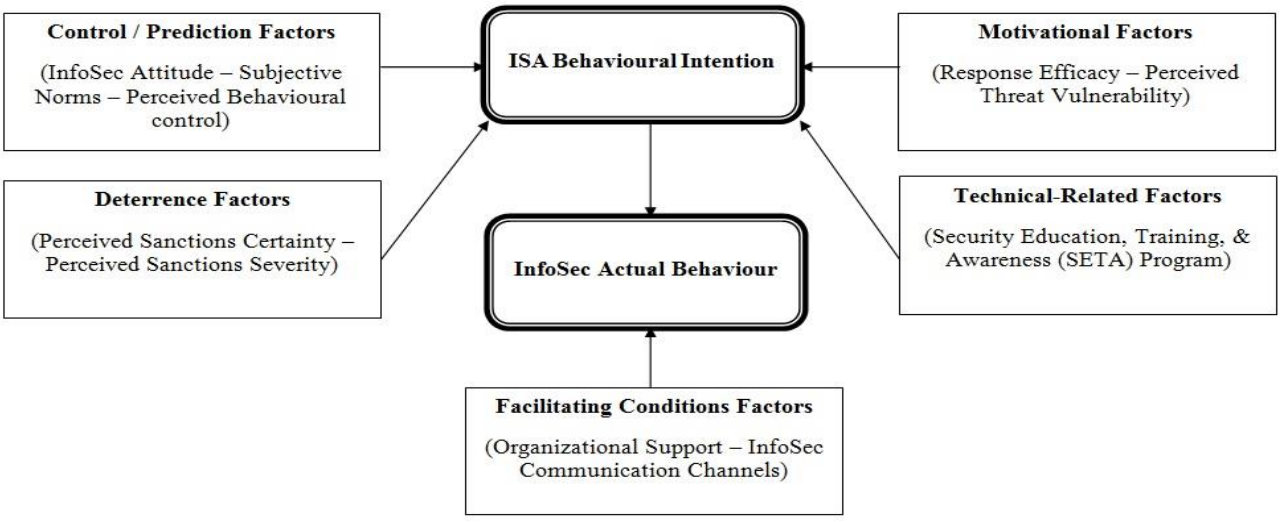


Fig. 1. Research Model.

A. Control and Prediction Factors

Many previous studies extracted prediction and control factors from TPB theory, and its constructs were employed in the field of information technology, proving their effectiveness by controlling employees' beliefs [3], [18], [27], [28]. According to Ong and Chong [29], some researchers have benefited from more helpful and practical recommendations due to citing TPB. Additionally, some studies [30-33] have applied the TPB to predict ISP compliance, information security awareness, and knowledge sharing from an individual's behavioural perspective, making the TPB more applicable to describing how employees participate in ISA activities. Hence, avoiding and mitigating information security breaches. Commonly, human interactions influence a person's beliefs and emotions, thoughts, feelings, behaviours, and actions. TPB contends that attitudes, subjective norms, and perceived behavioural control all impact intentions, which are the foundation of motivation to do behaviour [34]. Several studies have explored the association between attitude and intention [18], [28]. Attitude determines intention, according to TPB [34]. The study intends to utilise InfoSec attitude to represent an individual's acceptance or rejection of an idea. Thus, the employee's positive InfoSec attitude towards ISA reflects his/her intention. Conversely, negative InfoSec attitudes will reduce his/her intention. Consequently, an employee who positively believes in ISA is willing to engage in ISA activities and vice versa. Thus, it is hypothesised that:

H1: Employees' InfoSec attitudes toward ISA have a positive impact on their intention to participate in ISA activities.

Subjective norms are the perceived societal constraints exerted on a person to engage in or abstain from a specific behaviour [32]. Under this social pressure, a person acquires a set of norms, values, beliefs, and motives from significant individuals such as executives, managers, and co-workers [35]. When vital individuals exercise positive pressure on the employee in the context of ISA, this positively impacts the employee's intentions [33]. Thus, it is hypothesised that:

H2: Subjective norms towards ISA engagement positively affect employees' behavioural intentions.

Perceived behavioural control is a critical component of TPB [3], which refers to an individual's sense of how easy or difficult a task or action is to accomplish. Research in the information security arena has shown that perceived behavioural control has a significant effect on behavioural intentions [15], [28], [33]. In the current study, perceived behavioural control refers to the perception that adopting information security awareness is not difficult and has a positive impact. Thus, it is hypothesised that:

H3: Perceived behavioural control towards ISA has a positive impact on the behavioural intentions of employees.

B. Motivational Factors

According to the relevant literature, the PMT model is considered one of the best theories for predicting and motivating a person's intention to take preventive steps [17], [36]. The PMT theory was developed by Rogers [37] to understand fear appeals and predict suitable responses for personal protection when faced with a threat. When a person learns about potential threats, he or she becomes more aware of the risks to which he or she may be exposed. Threat appraisal and a coping appraisal are two primary constructs in PMT. The act of determining the intensity and sensitivity of danger is referred to as threat appraisal. While evaluating the success of protective measures and the perceived self-efficacy of the person under threat is referred to as coping appraisal. Empirical studies [2], [7], [14], [38] have shown the efficacy of PMT in implementing adherence and compliance to security standards and policies among an organisation's employees. Because these were the components found to have a positive influence in the literature related to the topic, this study used one factor from threat appraisal constructs: perceived vulnerability, and one from coping appraisal constructs: response efficacy. The perceived vulnerability relates to a person's appraisal of a potentially harmful circumstance and whether or not he or she is at risk [17]. Employees who perceive a high level of vulnerability in their organisation's information systems are more likely to take preventative measures. According to previous study findings [38], employees' perceived vulnerability in a cyber-attack incident encourages them to engage in preventive measures. As a result, it stands to reason that people who believe they are not vulnerable to security risks lack appropriate security knowledge and often fail to comply with workplace security policies. On the other hand, people who believe they are more exposed to security risks are more likely to engage in ISA activities and participate in preventative activities [39]. Thus, it is hypothesised that:

H4: Perceived vulnerability toward ISA has a positive impact on the behavioural intentions of employees.

Response efficacy relates to an individual's belief that adopting or implementing a certain preventative measure is the best method to reduce security risks [40]. When a person is persuaded of the utility of a risk-reduction mechanism, he or she will almost certainly adopt risk-reduction behaviour. However, if the person is not persuaded, he or she will not adopt it [17], [36], [41]. As a result, if employees think ISA gives them enough information and awareness to keep information security breaches and risks from happening, they are more likely to be motivated to participate in ISA activities. Thus, it is hypothesised that:

H5: Response efficacy towards ISA has a positive impact on the behavioural intentions of employees.

C. Deterrence Factors

The earliest version of the deterrence theory was created by the philosophers Cesare Beccaria and Jeremy Bentham, based on the assumption that individuals seek to maximise pleasant outcomes, such as rewards, and avoid painful ones, such as penalties [42]. GDT has been chiefly used in criminology to minimise deviant behaviour in people. In recent decades, it has been successfully and efficiently used for information technology as well as preventative information security [15], [19], [20], [27], [43]. In GDT, the deterrence model is built on three core constructs: certainty of sanctions, the severity of sanctions, and celerity of sanctions. Such determinants impact people's attitudes toward preventing activities that are regarded as undesirable in society. The constructs' of GDT: perceived

certainty of sanctions and perceived severity of sanctions are included in the study model due to their positive influence in the relevant literature [20], [43]. Perceived certainty of sanctions refers to a person's belief that the authorities are likely to detect delinquent behaviour. In contrast, the perceived severity of sanctions refers to the person's belief that s/he would be punished seriously if deviant behaviour is proven [3], [45]. When employees who break information security policies understand the consequences of their actions, they are more likely to participate in ISA activities and thus change their behaviour. Thus, it is hypothesised that:

H6: Perceived certainty of sanctions towards ISA has a positive impact on the behavioural intentions of employees.

H7: Perceived severity of sanctions towards ISA has a positive impact on the behavioural intentions of employees.

D. Technical-related Factors

Previous research has looked at the role of technical-related factors in improving ISA among users. Studies have a wide variety of interests in awareness-related variables that may not be within the vast area of education, training, and awareness. For example, the integrated model of Ramalingam et al. [46] used "Threat Awareness", "Password Awareness", and "Content Awareness"; Hanus and Wu [40] used "Threat Awareness" and "Countermeasure Awareness". Furthermore, Han [47] used "Security Technology Awareness"; Mamonov and Benbunan-Fich [48] used "Threat Awareness"; Khan and AlShare [49] used "information security policy scope". Moreover, Koohang et al. [50] used "Security Issues Awareness" and "Security Policy Awareness"; and Hwang et al. [51] used two separate constructs: "Security Policy" and "Security Education". According to Yaokumah et al. [52], security education benefits employees by improving their awareness of the organization's security environment, policies, and regulations. Effective training programmes may teach employees how to make secure information security choices. Staff security awareness programs may aid in the improvement of their security behaviour. Security education, training, and awareness (SETA) programs are educational and training programs designed to increase employees' knowledge of information security. These programs foster continued interest in rules and guidelines, risks, and the skills required to perform information systems security activities [21]. Consequently, rather than using the limited constructs of security awareness, the study prefers to use SETA as a construct with its complete and comprehensive concept of education, training, and awareness as compared to the limited constructs of security awareness. Employees may think they have the requisite knowledge and abilities to handle security issues in the workplace if they perceive SETA as effective. It stands to reason that employees with sufficient training are better equipped with skills and knowledge regarding security regulations and countermeasures. As a result, their behaviour will improve in order to comply with security policies. Hence, it is hypothesised that:

H8: SETA programs have a positive impact on the behavioural intentions of employees.

E. Behavioural Intention

One of the most significant constructs in TPB is the intention, which refers to the state of mind of a person in which the planning and forethought are to achieve a particular behaviour [3], [33]. According to the relevant studies, an individual's desire to achieve a goal that satisfies him or her yields an intention to participate in behaviour that encourages that goal. Bélanger et al. [14] and Thompson et al. [39] demonstrated that early conformity behavioural intention significantly predicts early conformity actual behaviour. In an attempt to predict the first adoption of information security behaviours, Ofori et al. [19] and Shropshire et al. [53] demonstrated a significant correlation between intention and actual behaviour. Although positive behavioural intentions toward a specific behaviour may ensure that the actual behaviour is achieved [51], intention alone may not adequately determine actual behaviour if explanatory power is not obtained by investigating both. Thus, it is hypothesised that:

H9: Employees' behavioural intentions towards ISA positively affect their adoption of InfoSec actual behaviour.

F. Facilitating Conditions Factors

External factors termed "facilitating conditions" (FC) are external factors outside the original theories. FCs are influential determinants that, along with other factors, promote a particular behaviour and are used to promote behavioural intention or actual behaviour to adopt technology [66]. These factors are included in the study's model to make an action easy to do. The study's model contains two constructs of facilitating conditions: organisational support [33] and InfoSec communication channels to promote employees' behaviour according to information security regulations. Organisational support indicates to employees; global beliefs about how an institution recognises and appreciates the employees' contributions and cares for their well-being. As Ofori et al. [19]; Khan and AlShare [49]; and Safa et al. [22] point out, organisations that show a commitment to their employees' well-being are better capable of protecting their assets through knowledge sharing and collaboration. Thus, it is hypothesised that:

H10: Organizational support towards employees facilitates their InfoSec's actual behaviour in accordance with information security policies.

Employee perception of the value of information and an organization's information security communication all contribute to the improvement of ISA through increasing knowledge of the importance of information security [1]. Moreover, employee communication channels regarding information security may reduce ambiguity and increase the frequency and usefulness of cross-functional communication, hence improving an individual's behaviour formation efficiency. Without formal communication channels, attitudes that violate safety norms would spread rapidly and prevent adopting correct ones [15]. According to Bélanger et al. [14], institutions may increase employee knowledge and awareness through targeted communications about the new need and justification for the recommended measures and security-related training. In terms of communication channels, this study asserts that effective communication amongst staff about

all information security concerns and issues may help reduce human vulnerabilities associated with having adequate expertise to comply with applicable laws and regulations. Thus, good communication can help employees learn new skills, make better decisions, report incidents, and clear up misconceptions about information security [13]. Hence, it is hypothesised that:

H11: InfoSec communication channels positively affect employees' InfoSec actual behaviour.

III. Methodology

This study aims to demonstrate how public organisations can manage the human component and increase their ISA by examining factors such as prediction, control, motivation, deterrence, technical-related, and facilitating conditions for the adoption of InfoSec behaviour and reducing the risk of information security breaches. The success factors were designed to maximise employees' ISA by relying on constructs from TPB, PMT, and GDT, as well as three external factors. Hence, this study methodology adheres to a positivist philosophy, which involves identifying essential relationships relating to the phenomenon (in this instance, the adoption of InfoSec behaviour); it also adheres to a quantitative approach, which is implemented via the distribution of a questionnaire. Expert feedback was used to develop the research model. Quantitative approaches were also used to enhance the model. Because the research is aimed at public sector units' employees, data was gathered from public government organisations in the Sultanate of Oman. A questionnaire with a 5-point Likert scale was used to gather data.

A. Instrument Development and Data Collection

After consulting questions from relevant past research, the questions in the present study's questionnaire were constructed to correspond to the framework and constructs. The questionnaire was divided into two sections: the first included six questions on the participants' demographics, and the second included questions about the proposed model's variables, for a total of 71 questions. In the final form of the questionnaire, each component was addressed with different questions with various options ranging from strongly disagree to strongly agreed (Using a Likert scale of 5-points). Before distribution, a pilot study with 100 respondents was conducted to ensure the reliability of the questionnaire's items [16] and to determine whether the questionnaire's questions were appropriate, intelligible, and subject to a single interpretation. The current study's data collection started in the first week of January 2022 and was finished by the end of February 2022 (Over approximately seven weeks). After describing the purpose of the study to the participants, we asked them to answer the questionnaire based on their knowledge and experience. Their consent was necessary for the researchers. They were given the questionnaire after confirming their consent to participate in this research. Participants were informed that their responses would be used exclusively for statistical and scholarly reasons and would be kept private. The study used stratified random sampling, which divides a target population into smaller subgroups called "strata". Random samples are drawn from these groups based on how much of the target population they make up.

B. Participants' Demographic Characteristics

The Sultanate of Oman's public sector employs 170,104 employees [54], making it one of the major sectors in the country. According to Krejcie and Morgan's equation [55], a sample of at least 384 participants is necessary for this study's intended population. Employees in the public sector were given 480 questionnaires, of which 415 were returned. The overall response rate was 86%, with 24 outliers. An overall response rate of 81% was obtained from the 391 validated responses. The remaining responses were discarded due to their repetitive answers or incompleteness. As shown in Table I, males comprised 248 (63.4%) of the total participants. The group over 40 years had the highest frequency of respondents' age, with 119 (30.4%), followed by 31—35 years of age, with 112 (28.6%). A bachelor's degree was the most often mentioned qualification among respondents (174; 44.5%). The most frequently occurring occupation among respondents (163; 34.8%) was "Employee", followed by "Technician" (70; 17.9%). The group with more than ten years of experience had the highest frequency of responders with more than ten years of experience (208; 53.2%). Most respondents belonged to educational or service-related institutions, with 128 (32.7%) and 101 (25.78%), respectively.

TABLE I. Demographic Characteristics

Variables		Frequency	%
Gender	Male	248	63.4
	Female	143	36.6
Age (years)	25 or Less	20	5.1
	26 – 30	33	8.4
	31 – 35	112	28.6
	36 - 40	107	27.4
	Above 40	119	30.4
Education	Diploma	52	13.3
	High Diploma	82	21
	Bachelor	174	44.5
	Master	71	18.2
	Doctorate	12	3.1
Employment Situation	Employee	136	34.8
	Specialist	53	13.6
	Technician	70	17.9
	Chief-Employee	47	12
	Manager	30	7.7
	Other	55	14.1
Experiences	1 - 2	26	6.6
	3 - 5	48	12.3
	6 - 10	109	27.9
	Above 10	208	53.2
Organization	Education	128	32.7
	Health	75	19.2
	Service	101	25.8
	Other	87	22.3

IV. RESULTS

A structured questionnaire adapted from prior studies was used to address the proposed conceptual model, which was then translated from English into Arabic and distributed to the target population. It was because the language of the survey had changed from one language to another that both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were used. SEM with AMOS version 24 was also used to see if the research hypotheses were accepted or rejected. Both measurement (MM) and structural (SM) models were developed in the study to model talent variables. They are two essential components used in the SEM to verify the study model's validity and reliability. The measurement model examines the relationship between latent constructs and their items to see if these indicators accurately measure the relevant talent construct. This step must be done before fitting the MM to the data to check the reliability and validity of the factor's items. In contrast, the structural model examines the relationships between one latent construct and other latent constructs [56].

A. Measurement Model Testing

SEM is a suitable approach for assessing data and estimating associations between constructs by accepting or rejecting formulated hypotheses to investigate the relationships between constructs in the study's model. SEM has many advantages, like isolating errors and estimating regression between latent constructs. Skewness and kurtosis tests were used to test the data distribution's normal state (normality); the study followed Hair et al. [57]'s recommendations and utilised a critical cut-off value of ±2.58. The results indicated that the skewness and kurtosis values for each model's variables were within the specified limits, indicating that the distribution is normal. For determining the suitability of factor analysis, Bartlett's test of sphericity (significant at $p<0.001$) and the Kaiser-Mayer-Olkin test (KMO) (values ranging between 0 and 1) were conducted. [58]. The KMO test score was 0.919, with a minimum suggested score of 0.50 and values greater than 0.9 were deemed excellent. The Chi-square statistic was significant (14100.952). The results of the KMO and Bartlett's tests are shown in Table II.

In line with relevant existing literature, the model for this study was constructed by incorporating the most successful parameters from three psychological theories and three external factors. Consequently, confirmatory factor analysis (CFA) is an important second step in evaluating whether the measured determinants align with our interpretation of the proposed model [59]. Furthermore, to develop the best potential measurement model, every item or latent variable that was not a good match (not fit) must be excluded [56]. The most frequent model-fit measures, according to Bollen [60], are the chi-square test (χ^2), comparative fix index (CFI), incremental fix index (IFI), Tucker-Lewis index (TLI), and root mean square error of approximation (RMSEA). Hence, in this study, these measures and the p-value were utilized as a goodness of fit indices to analyze the exogenous and endogenous variables. As a finding, χ^2 = (2558.954), degrees of freedom = (1346), ratio-χ^2/df = (1.901) less than 5, CFI = (.910), IFI = (.911), TLI = (.901), and RMSEA = (.048) less than 0.080, indicating that the measurement model was a good match (fit) with the data gathered [57]. Furthermore, the Root Mean Square Residual (RMR) = (0.037), less than 0.10. According to Table III, all model-fit indices surpassed the indicated acceptable thresholds.

The study used CFA to calculate the factor loading of the measurement variables to estimate the convergent viability. According to Hair et al. [57], if the loading factor of the indicators is more than 0.50 and the sample size is 300 or above, the loading factor shows an acceptable level of convergent validity. As a result, we removed indicators from the study's model with a factor loading of less than 0.50. Due to lower factor loadings (less than 0.50) or cross-loadings, the indicators SN1, SN5, and SN6 in subjective norms, PBC1 in perceived behavioural control, PV5 in perceived threat vulnerability, RE1, RE2 in response efficacy, PCOS1 in perceived sanctions certainty, PSOS1, PSOS2, and PSOS3 in perceived sanctions severity, SETA2 and SETA7 in security education, training, and awareness, BI7 in behavioural intention, OS1 in organisational support, and COM5 in InfoSec communication channels were eliminated from the proposed model. Internal consistency in the measuring of model variables is provided through reliability measurement. A questionnaire's reliability (Cronbach's alpha) is thought to be accepted when it is more than 0.6 [61], and when it is above 0.7, it is indicated to be composite reliability [56]. The two kinds of reliability testing were used in this analysis. Cronbach's alpha scores vary from 0.807 to 0.908, while composite reliability scores range from 0.814 to 0.901. As a result, the reliability and composite reliability values for the entire model's variables were more than 0.7. Table IV provides an overview of the statistical measurements.

TABLE II. THE KMO AND BARTLETT'S TEST RESULTS

Measurement of Sampling Adequacy: Kaiser-Meyer-Olkin	0.919	
Bartlett's Sphericity Test	**Approx. Chi –Square**	14100.952
	Df	1485
	Sig	.000

TABLE III. MM AND SM FIT INDICES

Fit Index	Cut-off Points	MM	SM
X^2	-	2558.954	3378.335
d.f	-	1346	1393
Ratio (X^2/d.f)	<5	1.901	2.425
CFI	>0.90	.910	.853
IFI	>0.90	.911	.854
TLI	>0.90	.901	.850
RMSEA	<0.08	.048	.060
RMR	<0.10	0.037	-

TABLE IV. THE VARIABLES, MEASURES, AND THEIR DESCRIPTIVE STATISTICS

Variables	Items	Measures	Factor Loading	AVE	Alpha	CR
InfoSec Attitude	ATT1	Information security awareness is necessary.	0.773	0.514	0.870	0.862
	ATT2	Information security awareness is beneficial.	0.877			
	ATT3	Practicing information security awareness activities is useful.	0.754			
	ATT4	I believe that information security awareness is a useful behavioural tool to safeguard the organization's information assets.	0.633			
	ATT5	My information security awareness has a positive effect on mitigating the risk of information security breaches.	0.624			
	ATT6	Information security awareness is a wise approach that decreases the risk of information security incidents.	0.624			
Subjective Norms	SN2	My colleagues think that I should have information security awareness to protect organizational information assets.	0.685	0.545	0.864	0.781
	SN3	My friends in my office encourage me to understand information security policies.	0.699			
	SN4	The head of the department thinks that information security awareness is a value culture	0.822			
Perceived Behavioural Control	PBC2	I have the necessary awareness about information security to share with the other employees.	0.766	0.636	0.873	0.875
	PBC3	I have the ability to adopt information security awareness to mitigate the risk of information security breaches.	0.804			
	PBC4	Information security awareness adoption is an easy and enjoyable task for me.	0.803			
	PBC5	I have enough knowledge to behave safely in terms of information security.	0.817			
Response Efficacy	RE3	At my work, efforts to ensure the safety of my confidential information are effective.	0.668	0.661	0.858	0.852
	RE4	The preventative measures available to me to stop people from gaining access to my organization’s information are adequate.	0.880			
	RE5	The preventative measures available to me to prevent people from damaging my information system at work are adequate.	0.873			
Perceived Threat Vulnerability	PV1	I know my organization could be vulnerable to security breaches if I don’t adhere to its information security policy.	0.755	0.524	0.812	0.814
	PV2	I could fall victim to a malicious attack if I fail to comply with my organization’s information security policy.	0.727			
	PV3	I believe that trying to protect my organization’s information will reduce illegal access to it.	0.673			
	PV4	My organization’s data and resources may be compromised if I don’t pay adequate attention to guidelines.	0.737			
Behavioural Intention	BI1	I am willing to practice my information security awareness because of its potential to reduce the risks.	0.674	0.501	0.872	0.857
	BI2	I will share my information security awareness with my colleagues to comply with security policies.	0.734			
	BI3	I intend to help my colleagues to increase their awareness of information security	0.776			
	BI4	I intend to collaborate with other staff to decrease insider threats in my organization.	0.699			
	BI5	I will inform the other staff about new methods and software that can reduce the risk of information security.	0.686			
	BI6	I will share the report on information security incidents with others, in order to reduce the risk.	0.665			
InfoSec Actual Behaviour	AB1	I frequently practice my experience about information security with my colleagues.	0.663	0.505	0.908	0.901
	AB2	I practice my information security knowledge with my colleagues.	0.653			
	AB3	I frequently share my expertise from my information security training with my colleagues.	0.680			
	AB4	I frequently talk with others about information security incidents and their solutions in our meetings.	0.709			
	AB5	I avoid mistakes in the domain of information security.	0.783			

	AB6	I always mitigate information security threats.	0.785			
	AB7	I think about the consequences of my behaviour before any action.	0.666			
	AB8	I am careful about my behaviour in the domain of information security.	0.642			
	AB9	I frequently assess my information security behaviour to improve it.	0.750			
Perceived Certainty of Sanctions	PCOS2	I believe that if I violate the confidentiality of information, the management will realise it.	0.769	0.640	0.866	0.875
	PCOS3	If I violated the organization's security policies, I would probably be caught.	0.624			
	PCOS4	I believe that if I transfer organisational information outside, the organisation will find out about my violation.	0.907			
	PCOS5	I believe that if I sell organisational information, my organisation will discover it.	0.870			
Perceived Severity of Sanctions	PSOS4	I deserve punishment if I violate the confidentiality of organisational information.	0.782	0.605	0.807	0.820
	PSOS5	I think punishment will be high if I sell or transfer organisational information outside.	0.854			
	PSOS6	I think receiving sanctions because of my information security misconduct will negatively influence my career development.	0.688			
Organizational Support	OS2	The organisation cares about my information security awareness level.	0.839	0.654	0.863	0.883
	OS3	The management appreciates employees for their information security awareness.	0.805			
	OS4	The management awards employees for their compliance with information security policies.	0.766			
	OS5	The management encourages employees to participate in information security awareness engagement.	0.824			
InfoSec Communication Channels	COM1	We have communication channels established for employees to report information security suspected improprieties.	0.755	0.669	0.875	0.889
	COM2	The management communicates employees' security duties and control responsibilities in an effective manner.	0.922			
	COM3	Communication flows across the organisation adequately (e.g., from department to department) to enable employees to discharge their responsibilities securely and efficiently.	0.753			
	COM4	I feel as though I am a part of the information security decision-making process within my organization.	0.727			
Security Education, Training and Awareness	SETA1	My organization gives employees training to help them become more aware of information system security issues.	0.610	0.568	0.866	0.866
	SETA3	SETA increases my knowledge of security issues.	0.752			
	SETA4	SETA motivates the learners to integrate the security knowledge taught.	0.855			
	SETA5	My organisation provides employees with appropriate security education before giving them authorised access to the institution's network.	0.844			
	SETA6	My organization utilizes various communication methods in order to improve the information security awareness of employees.	0.679			

TABLE V. CORRELATION ANALYSIS AND DISCRIMINANT VALIDITY

	AVE	MSV	PV	PBC	OS	PCOS	COM	SETA	ATT	SN	AB	PSOS	RE	BI
PV	0.524	0.309	**0.724**											
PBC	0.636	0.530	0.390	**0.798**										
OS	0.654	0.605	0.343	0.509	**0.809**									
PCOS	0.640	0.411	0.416	0.430	0.641	**0.800**								
COM	0.669	0.605	0.368	0.490	0.778	0.617	**0.818**							
SETA	0.568	0.378	0.372	0.615	0.610	0.485	0.562	**0.754**						
ATT	0.514	0.131	0.362	0.174	-0.077	0.017	0.023	0.167	**0.717**					
SN	0.545	0.041	-0.026	-0.203	-0.194	-0.201	-0.197	-0.174	0.058	**0.738**				
AB	0.505	0.500	0.505	0.728	0.643	0.475	0.575	0.588	0.117	-0.146	**0.710**			
PSOS	0.605	0.279	0.528	0.289	0.383	0.494	0.385	0.374	0.268	-0.066	0.523	**0.778**		
RE	0.661	0.444	0.385	0.652	0.657	0.515	0.618	0.586	0.005	-0.078	0.666	0.290	**0.813**	
BI	0.501	0.462	0.556	0.475	0.528	0.344	0.379	0.460	0.330	0.044	0.680	0.485	0.442	**0.707**

Note: PV = Perceived Threat Vulnerability, PBC = Perceived Behavioural Control, OS = Organizational Support, PCOS = Perceived Certainty of Sanctions, COM = InfoSec Communication Channels, SETA = Security Education, Training and Awareness, ATT = InfoSec Attitude, SN = Subjective Norms, AB = InfoSec Actual Behaviour, PSOS = Perceived Severity of Sanctions, RE = Response Efficacy, BI = Behavioural Intention.

Discriminant validity is realized when a construct is remarkably different from the other constructs since there is no association between constructs that do not relate to each other [57]. The square root of the AVE was more significant than the correlations between the construct and the other model's constructs, which varied between 0.017 and 0.778 for the given model. Moreover, the maximum shared squared variance (MSV) was smaller than the AVE. Thus, the discriminant validity verification supported all of the model's constructs. Table V displays the matrices of correlation between various latent variables.

B. Structural Model Testing

In this study, the same set of fit indices is used to analyse the structural model. As indicated in Table III, all fit indices were within the acceptable ranges: χ^2 = (3378.335), degrees of freedom= (1393), ratio-χ^2/df= (2.425), RMSEA = (.060), with the exception of CFI = (.853), IFI = (.854), and TLI = (.850). However, another method of evaluating the values derived from the CFI, IFI, and TLI indices should be considered. According to Bentler and Bonett [62] and Sharma et al. [63], the TLI cut-off point is continually shifting. Since there is no globally approved measuring standard, a TLI value between 0.80 and 0.90 may be considered a moderate or acceptable fit. Bentler [64] believed that CFI indicates a good fit when it equals or surpasses 0.90, while values larger than 0.80 and reaching 0.90 suggest a generally adequate fit, and Bollen [60] made the same suggestion for IFI index values. Moreover, Schumacher and Lomax [65] state that if the IFI, CFI, and ITL values are greater than 0.90, they are considered excellent fits, but they may also be considered moderate if the values are between 0.85 and 0.90. As a result of the above, we believe that the model is both appropriate and a good match for the data, as the parsimonious index provides the most accurate measurement (RMSEA= .060).

TABLE VI. STRUCTURAL MODEL CAUSAL PATHS

Paths	Standardized estimate (β)	P-value	Result
ATT → BI	0.138	0.009	Supported
SN → BI	0.146	0.020	Supported
PBC → BI	0.300	0.000	Supported
PV → BI	0.311	0.000	Supported
RE → BI	0.148	0.045	Supported
PCOS → BI	-0.107	0.106	Rejected
PSOS → BI	0.276	0.000	Supported
SETA → BI	0.139	0.000	Supported
BI → AB	0.582	0.000	Supported
OS → AB	0.262	0.001	Supported
COM → AB	0.187	0.015	Supported

Note: ATT = InfoSec Attitude, SN = Subjective Norms, PBC = Perceived Behavioural Control, RE = Response Efficacy, PV = Perceived Threat Vulnerability, PCOS = Perceived Certainty of Sanctions, PSOS = Perceived Severity of Sanctions, SETA = Security Education, Training and Awareness, BI = Behavioural Intention, OS = Organizational Support, COM = InfoSec Communication Channels, AB = InfoSec Actual Behaviour.

The findings of the causal paths are shown in Table VI. Employees' ISA behavioural intention was significantly influenced by InfoSec attitude (β=0.137, p=0.009), subjective norms (β=0.107, p=0.048), perceived behavioural control (β=0.296, p=0.000), response efficacy (β=0.148, p=0.018), perceived threat vulnerability (β=0.297, p=0.000), perceived sanctions severity (β=0.274, p=0.000), and security education, training, and awareness (β=0.139, p=0.000). On the other hand, the impact of perceived sanctions certainty on employees' ISA behavioural intentions was insignificant. As a result, H6 is rejected. Finally, the results demonstrated that ISA behavioural intention (β=0.584, p=0.000), InfoSec communication channels (β=0.188, p=0.015), and organizational support (β=0.262, p=0.001) all had a significant impact on InfoSec actual behaviour adoption.

V. COMPARATIVE EVALUATION OF THE STUDY MODEL

The study's significance derives from the inclusion of control, prediction, motivation, and deterrence approaches, all resulting from three main theories: TPB, PMT, and GDT. This study investigated whether the TPB affected intentions in information security behaviour adoption among public organisation employees and revealed that the TPB has a good to excellent effect, supporting results of previous studies [3], [18], [27], [28] and contradicting the findings of Rajab and Eydgahi's [2] study. The presented factors encourage institutions' employees to engage in ISA activities and, consequently, InfoSec behaviour adoption. The results of the InfoSec attitude analysis indicated that employees who expect advantages from ISA activities are more likely to adopt InfoSec behaviours consistent with their understanding of ISA. As a consequence of our analysis of subjective norms, we can assume that employees get cooperation about their engagement in ISA activities from their managers, supervisors, and co-workers. The present case demonstrates the significance of perceived behavioural control, which indicates that controlling perceptions may impact employees' intentions, allowing ISA activities to effectively engage in a suitable work environment. Because PMT is a practical framework for estimating an employee's intention to take preventive measures, some studies indicate that perceived vulnerability [17], [39] and response efficacy [40], [41] related to information security have a significant impact on information security policy compliance. This study found that almost all of their findings align with these findings. The study also found PMT to be among the best theoretical frameworks for explaining ISA intentions toward InfoSec behaviour adoption, which is consistent with previous results [17], [36], [39]. The purpose of GDT constructs is to treat employee criminal behaviour. The target of applying sanctions is to prevent or eliminate undesirable employee conduct. The imposition of sanctions helps to alter the behaviour of uncooperative staff to some degree [44] and raises awareness of illegal behaviour among other employees when penalties are implemented. As proven by prior studies [19], [27], [43], there is a significant positive relationship between the severity of sanctions and compliance with information security policies. While the results of this study confirm the findings of earlier research on the sanctions' severity and InfoSec's behaviour through ISA intention, they also suggest that as the likelihood of sanction severity rises, employees' intentions for InfoSec behaviour rise as well. Jaeger et al. [27] discovered that the sanctions' certainty did not affect the variance in information security policy compliance, which

confirms the results of this study. The study targeted public sector employees as a possible explanation for this non-significant relationship. This sector most certainly differs from the private sector in several ways; for example, employees in this sector work in a more stable environment, which may lower their motivation compared to private sector employees. Employees in the public sector might need more powerful ways to get them excited, such as recognition and responsibility.

The study's findings reveal that SETA programs strongly affect public institution employees' ISA intentions towards InfoSec behaviour. According to prior studies, SETA programs motivate employees to follow information security policies and procedures [20], [21]. When public sector employees get appropriate SETA programs, they will gain an essential awareness of security knowledge and abilities. They will also be able to show their commitment to the information security policy through their behaviour. Accordingly, they will be one of the most effective defence lines in safeguarding information assets and professionally responding to risks and attacks. Furthermore, the results showed that a positive ISA intention toward adopting InfoSec behaviour, organisational support, and InfoSec communication channels affected employees' adoption of InfoSec behaviour. The statistical analysis and the literature review show that the proposed model is both sound and efficient. A model for adopting information security behaviours in public organisations was contributed by determining the success factors that would influence the intentions of public sector employees to engage in ISA activities and adopt best behavioural practices. It is expected that the results of this research will be used by content development consultants to improve and enhance ISA materials and by SETA program developers and designers to prepare and design ISA and best practices programs and initiatives. The proposed model concentrates on the two aspects of behavioural intentions and actual behaviour to add to existing knowledge on ISA and best practices. In addition, it includes the facilitating conditions that positively influence employees' ISA (i.e., Organisational support and InfoSec communication channels) to enhance and correct actual behaviour in the process of ISA.

This study is one of the studies that envision increasing employee awareness and understanding of information security and reducing breaches through a combination of factors. This aggregation creates a new perspective that helps public institutions more effectively manage human ISA. We believe this research adds to this field's existing body of knowledge.

VI. Conclusion, Limitations and Future Work

The fast growth of information technology has made it simpler, more accurate, and more efficient to carry out organisational functions. Nevertheless, there is still a gap between how far technology has advanced and how much employees are aware of it, making it challenging for public institutions to preserve their assets. A lack of users' ISA causes many security risks and challenges. Through this study, we seek to strengthen and broaden research on the challenges of ISA in organisations by leveraging success factors extracted from three theories established on the principles of control, prediction, motivation, and deterrence. Public institutions may influence their employees' intentions to align with desired information security behaviour by employing control and prediction factors. Employees are also encouraged by motivational factors to practice security countermeasures and continuously maintain their knowledge and skills. Deterrence factors contribute to the control of criminal wrongdoing and, through them, can spread security awareness via understanding criminal behaviour. Usually, there are two aspects to SETA programs: the fundamental part and the institutional-specific. The fundamental part of all SETA programs is to determine and monitor the critical human threats and risks and employee behaviours linked to those threats and risks. The institutional-specific part is designed to address the requirements of employees and the institution. The institution's recognised risks and behaviours should influence the awareness efforts. Employees must be provided with these programs consistently. They must also be consistently evaluated. Furthermore, the research model has been expanded to include facilitating conditions that help make sure that actual InfoSec behaviour is in line with information security regulations and policies.

To further extend this study, it is necessary to identify the determinants that influence employees' engagement in ISA activities, their difficulties and obstacles, and their perspectives on them. Future studies might look at the implementation of ISA through alternative models and theories and extend the technical, organisational, environmental, and individual factors. Additionally, interviews and group discussions might be conducted to ascertain any underlying reasons for the lack of ISA, particularly in public institutions. Among the study's limitations is that it focused only on public-sector units in Oman. Consequently, the findings do not accurately reflect the behaviour of other sectors, such as the private, industrial, and financial sectors, resulting in a lack of representation. Future studies can incorporate a diverse range of sectors into their study sample.

Acknowledgment

This publication has been funded under the industrial grant INDUSTRI/APSB/FTMK/2021/I00063. The authors would like to thank UTeM and Ask Pentest Sdn Bhd research group members for their support.

References

[1] K. Khando, S. Gao, S. Islam and A. Salman, "Enhancing employees information security awareness in private and public organisations: A systematic literature review", *Computers & Security*, vol. 106, p. 102267, 2021.

[2] M. Rajab and A. Eydgahi, "Evaluating the explanatory power of theoretical frameworks on intention to comply with information security policies in higher education", *Computers & Security*, vol. 80, pp. 211-223, 2019.

[3] N. S. Safa, C. Maple, S. Furnell, M. A. Azad, C. Perera, M. Dabbagh, and M. Sookhak, "Deterrence and prevention-based model to mitigate information security insider threats in organisations", *Future Generation Computer Systems*, vol. 97, pp. 587-597, 2019.

[4] C. Cybersecurity, "ENISA threat landscape report 2018 : 15 top cyber-threats and trends.", *Op.europa.eu*, 2022. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/6373c334-574d-11e9-a8ed-01aa75ed71a1/language-en. [Accessed: 03- May- 2022].

[5] N. Humaidi and V. Balakrishnan, "Leadership Styles and Information Security Compliance Behavior: The Mediator Effect of Information

Security Awareness", *International Journal of Information and Education Technology*, vol. 5, no. 4, pp. 311-318, 2015.

[6] T. Sommestad, "Work-related groups and information security policy compliance", *Information & Computer Security*, vol. 26, no. 5, pp. 533-550, 2018. Available: 10.1108/ics-08-2017-0054.

[7] L. Li, W. He, L. Xu, I. Ash, M. Anwar and X. Yuan, "Investigating the impact of cybersecurity policy awareness on employees' cybersecurity behavior", *International Journal of Information Management*, vol. 45, pp. 13-24, 2019.

[8] F. Haeussinger and J. Kranz, "Antecedents of employees' information security awareness - review, synthesis, and directions for future research", *AIS Electronic Library (AISeL)*, 2022. [Online]. Available: https://aisel.aisnet.org/ecis2017_rp/12/. [Accessed: 03- May- 2022].

[9] M. Siponen, M. Adam Mahmood and S. Pahnila, "Employees' adherence to information security policies: An exploratory field study", *Information & Management*, vol. 51, no. 2, pp. 217-224, 2014.

[10] R. Vaidya, "Cyber Security Breaches Survey 2019", *Department for Digital, Culture, Media and Sport*, 2019. [Online]. Available: https://drj.com/wp-content/uploads/2019/04/Cyber_Security_Breaches_Survey_2019_-_Main_Report.PDF. [Accessed: 03- Apr- 2022].

[11] L. Jaeger, "Information security awareness: literature review and integrative framework", in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[12] J. Abawajy, "User preference of cyber security awareness delivery methods", *Behaviour & Information Technology*, vol. 33, no. 3, pp. 237-248, 2014.

[13] M. Bada, A. Sasse, and J. R. C. Nurse, "Cyber security awareness campaigns: Why do they fail to change behaviour?" in *International Conference on Cyber Security for Sustainable Society*, 2019.

[14] F. Bélanger, S. Collignon, K. Enget and E. Negangard, "Determinants of early conformance with information security policies", *Information & Management*, vol. 54, no. 7, pp. 887-901, 2017.

[15] Y. Hong and S. Furnell, "Motivating Information Security Policy Compliance: Insights from Perceived Organizational Formalization", *Journal of Computer Information Systems*, vol. 62, no. 1, pp. 19-28, 2022.

[16] I. Al-Shanfari, W. Yassin, R. Abdullah, N. Al-Fahim and R. Ismail, "Introducing A Novel Integrated Model for the Adoption of Information Security Awareness through Control, Prediction, Motivation, and Deterrence Factors: A Pilot Study", *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 1, pp. 2991–3003, 2021.

[17] C. Howell, "Self-Protection in Cyberspace: Assessing the Processual Relationship Between Thoughtfully Reflective Decision Making, Protection Motivation Theory, Cyber Hygiene, and Victimization", Doctoral dissertation, University of South Florida, 2021.

[18] T. Grassegger and D. Nedbal, "The Role of Employees' Information Security Awareness on the Intention to Resist Social Engineering", *Procedia Computer Science*, vol. 181, pp. 59-66, 2021. Available: 10.1016/j.procs.2021.01.103.

[19] K. Ofori, H. Anyigba, G. Ampong, O. Omoregie, M. Nyamadi and E. Fianu, "Factors influencing information security policy compliance behavior", in *Research Anthology on Business Aspects of Cybersecurity*, 2022, pp. 213-232.

[20] K. Kuo, P. Talley and D. Lin, "Hospital Staff's Adherence to Information Security Policy: A Quest for the Antecedents of Deterrence Variables", *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, vol. 58, pp. 1-12, 2021.

[21] H. Kim, H. Choi and J. Han, "Leader power and employees' information security policy compliance", *Security Journal*, vol. 32, no. 4, pp. 391-409, 2019.

[22] N. Sohrabi Safa, C. Maple, T. Watson and S. Furnell, "Information security collaboration formation in organisations", *IET Information Security*, vol. 12, no. 3, pp. 238-245, 2018.

[23] P. Kuppusamy, G.N. Samy, N. Maarop, P. Magalingam, N. Kamaruddin, B. Shanmugam, and S. Perumal, "Systematic Literature Review of Information Security Compliance Behaviour Theories", *Journal of Physics: Conference Series*, vol. 1551, no. 1, p. 012005, 2020.

[24] M. Alassaf and A. Alkhalifah, "Exploring the Influence of Direct and Indirect Factors on Information Security Policy Compliance: A Systematic Literature Review", *IEEE Access*, vol. 9, pp. 162687-162705, 2021.

[25] T. Gundu and S. Flowerday, "Ignorance to Awareness: Towards an Information Security Awareness Process", *SAIEE Africa Research Journal*, vol. 104, no. 2, pp. 69-79, 2013.

[26] A. Burns, T. Roberts, C. Posey, R. Bennett and J. Courtney, "Intentions to Comply Versus Intentions to Protect: A VIE Theory Approach to Understanding the Influence of Insiders' Awareness of Organizational SETA Efforts", *Decision Sciences*, vol. 49, no. 6, pp. 1187-1228, 2017.

[27] L. Jaeger, A. Eckhardt and J. Kroenung, "The role of deterrability for the effect of multi-level sanctions on information security policy compliance: Results of a multigroup analysis", *Information & Management*, vol. 58, no. 3, p. 103318, 2021.

[28] Y. Hong and S. Furnell, "Organizational formalization and employee information security behavioral intentions based on an extended TPB model", in *2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, 2019, pp. 1-4.

[29] L. Ong and C. Chong, "Information Security Awareness: An Application of Psychological Factors–A Study in Malaysia", in *2014 International Conference on Computer, Communications and Information Technology (CCIT 2014)*, 2014, pp. 98-101.

[30] B. Khan, K. Alghathbar, S. Nabi and M. Khan, "Effectiveness of information security awareness methods based on psychological theories", *African Journal of Business Management*, vol. 5, no. 26, pp. 10862-10868, 2011.

[31] A. Ahlan, M. Lubis and A. Lubis, "Information Security Awareness at the Knowledge-Based Institution: Its Antecedents and Measures", *Procedia Computer Science*, vol. 72, pp. 361-373, 2015.

[32] N. Safa, M. Sookhak, R. Von Solms, S. Furnell, N. Ghani and T. Herawan, "Information security conscious care behaviour formation in organizations", *Computers & Security*, vol. 53, pp. 65-78, 2015. http:10.1016/j.cose.2015.05.012.

[33] N. Safa and R. Von Solms, "An information security knowledge sharing model in organizations", *Computers in Human Behavior*, vol. 57, pp. 442-451, 2016. Available: 10.1016/j.chb.2015.12.037.

[34] I. Ajzen, "The theory of planned behavior", *Organizational Behavior and Human Decision Processes*, vol. 50, no. 2, pp. 179-211, 1991.

[35] A. Onumo, I. Ullah-Awan and A. Cullen, "Assessing the Moderating Effect of Security Technologies on Employees Compliance with Cybersecurity Control Procedures", *ACM Transactions on Management Information Systems*, vol. 12, no. 2, pp. 1-29, 2021. Available: 10.1145/3424282.

[36] L. Li, L. Xu and W. He, "The effects of antecedents and mediating factors on cybersecurity protection behavior", *Computers in Human Behavior Reports*, vol. 5, p. 100165, 2021. Available: 10.1016/j.chbr.2021.100165.

[37] R. Rogers, "A Protection Motivation Theory of Fear Appeals and Attitude Change1", *The Journal of Psychology*, vol. 91, no. 1, pp. 93-114, 1975. Available: 10.1080/00223980.1975.9915803.

[38] S. Boss, D. Galletta, P. Lowry, G. Moody and P. Polak, "What Do Systems Users Have to Fear? Using Fear Appeals to Engender Threats and Fear that Motivate Protective Security Behaviors", *MIS Quarterly*, vol. 39, no. 4, pp. 837-864, 2015. Available: 10.25300/misq/2015/39.4.5.

[39] N. Thompson, T. McGill and X. Wang, "'Security begins at home': Determinants of home computer and mobile device security behavior", *Computers & Security*, vol. 70, pp. 376-391, 2017. Available: 10.1016/j.cose.2017.07.003.

[40] B. Hanus and Y. Wu, "Impact of Users' Security Awareness on Desktop Security Behavior: A Protection Motivation Theory Perspective", *Information Systems Management*, vol. 33, no. 1, pp. 2-16, 2015. Available: 10.1080/10580530.2015.1117842.

[41] M. Martens, R. De Wolf and L. De Marez, "Investigating and comparing the predictors of the intention towards taking security measures against malware, scams and cybercrime in general", *Computers in Human Behavior*, vol. 92, pp. 139-150, 2019. Available: 10.1016/j.chb.2018.11.002.

[42] M. Stafford, "Deterrence Theory: Crime", *International Encyclopedia of the Social & Behavioral Sciences*, pp. 255-259, 2015. Available: 10.1016/b978-0-08-097086-8.45005-1 [Accessed 4 May 2022].

[43] N. Ameen, A. Tarhini, M. Hussain Shah and N. Madichie, "Employees' behavioural intention to smartphone security: A gender-based, cross-national study", *Computers in Human Behavior*, vol. 104, p. 106184, 2020. Available: 10.1016/j.chb.2019.106184.

[44] B. Lebek, J. Uffen, M. Neumann, B. Hohler and M. Breitner, "Information security awareness and behavior: a theory-based literature review", *Management Research Review*, vol. 37, no. 12, pp. 1049-1092, 2014. Available: 10.1108/mrr-04-2013-0085.

[45] P. Ifinedo, "Critical Times for Organizations: What Should Be Done to Curb Workers' Noncompliance With IS Security Policy Guidelines?", *Information Systems Management*, vol. 33, no. 1, pp. 30-41, 2015. Available: 10.1080/10580530.2015.1117868.

[46] R. Ramalingam, R. Lakshminarayanan and S. Khan, "Information Security Awareness at Oman Educational Institutions : An Academic Prespective", *arXiv.org*, 2016. [Online]. Available: https://arxiv.org/abs/1605.05580. [Accessed: 13- Mar- 2022].

[47] B. Han, "User's Information Security Awareness in BYOD Programs: A Theoretical Model", in *Information Institute Conference*, 2017.

[48] S. Mamonov and R. Benbunan-Fich, "The impact of information security threat awareness on privacy-protective behaviors", *Computers in Human Behavior*, vol. 83, pp. 32-44, 2018. Available: 10.1016/j.chb.2018.01.028.

[49] H. Khan and K. AlShare, "Violators versus non-violators of information security measures in organizations—A study of distinguishing factors", *Journal of Organizational Computing and Electronic Commerce*, vol. 29, no. 1, pp. 4-23, 2019. Available: 10.1080/10919392.2019.1552743.

[50] A. Koohang, J. Anderson, J. Nord and J. Paliszkiewicz, "Building an awareness-centered information security policy compliance model", *Industrial Management & Data Systems*, vol. 120, no. 1, pp. 231-247, 2019. Available: 10.1108/imds-07-2019-0412.

[51] I. Hwang, R. Wakefield, S. Kim and T. Kim, "Security Awareness: The First Step in Information Security Compliance Behavior", *Journal of Computer Information Systems*, vol. 61, no. 4, pp. 345-356, 2021. Available: 10.1080/08874417.2019.1650676.

[52] W. Yaokumah, D. Walker and P. Kumah, "SETA and Security Behavior", *Journal of Global Information Management*, vol. 27, no. 2, pp. 102-121, 2019. Available: 10.4018/jgim.2019040106.

[53] J. Shropshire, M. Warkentin and S. Sharma, "Personality, attitudes, and intentions: Predicting initial adoption of information security behavior", *Computers & Security*, vol. 49, pp. 177-191, 2015. Available: 10.1016/j.cose.2015.01.002.

[54] MOL, "Annual Report in the Civil Service Sector", *Staff.mol.gov.om*, 2021. [Online]. Available: https://staff.mol.gov.om/DSMVD/CMS/WebSiteMediaAnnual/07102021%20075839%20%D8%B5_vgyrm04r440la3h1sy3ksqin20219202175835(stat%202020).pdf. [Accessed: 14- Jan- 2022].

[55] R. Krejcie and D. Morgan, "Determining Sample Size for Research Activities", *Educational and Psychological Measurement*, vol. 30, no. 3, pp. 607-610, 1970. Available: 10.1177/001316447003000308.

[56] J. Hair, *Multivariate data analysis*. Englewood Cliffs, NJ: Prentice Hall, 2010.

[57] J. Hair, C. Ringle and M. Sarstedt, "Partial Least Squares Structural Equation Modeling: Rigorous Applications, Better Results and Higher Acceptance", *Long Range Planning*, vol. 46, no. 1-2, pp. 1-12, 2013. Available: 10.1016/j.lrp.2013.01.001.

[58] H. Kaiser, "An index of factorial simplicity", *Psychometrika*, vol. 39, no. 1, pp. 31-36, 1974. Available: 10.1007/bf02291575.

[59] R. Ho, *Handbook of univariate and multivariate data analysis and interpretation with SPSS*, 1st ed. Boca Raton, Fla: Chapman & Hall/CRC, 2006.

[60] K. Bollen, "A New Incremental Fit Index for General Structural Equation Models", *Sociological Methods & Research*, vol. 17, no. 3, pp. 303-316, 1989. Available: 10.1177/0049124189017003004.

[61] U. Sekaran, "Towards a guide for novice research on research methodology: Review and proposed methods", *Journal of Cases of Information Technology*, vol. 8, no. 4, pp. 24-35, 2003.

[62] P. Bentler and D. Bonett, "Significance tests and goodness of fit in the analysis of covariance structures.", *Psychological Bulletin*, vol. 88, no. 3, pp. 588-606, 1980. Available: 10.1037/0033-2909.88.3.588.

[63] S. Sharma, S. Mukherjee, A. Kumar and W. Dillon, "A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models", *Journal of Business Research*, vol. 58, no. 7, pp. 935-943, 2005. Available: 10.1016/j.jbusres.2003.10.007.

[64] P. Bentler, "Comparative fit indexes in structural models.", *Psychological Bulletin*, vol. 107, no. 2, pp. 238-246, 1990. Available: 10.1037/0033-2909.107.2.238.

[65] R. Schumacker and R. Lomax, *A beginner's guide to structural equation modeling*. psychology press, 2004.

[66] H. Triandis, "Values, attitudes, and interpersonal behavior", in *Nebraska Symposium on Motivation*, 1979, pp. 195-259.

[67] I. Al-Shanfari, W. Yassin, R. Abdullah, "Identify of Factors Affecting Information Security Awareness and Weight Analysis Process", *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, no. 3, pp. 534-42, 2020.

Novel Oversampling Algorithm for Handling Imbalanced Data Classification

Novel Oversampling Algorithm

Anjali S. More, Dipti P. Rana
Department of Computer Science and Engineering
Sardar Vallabhbhai National Institute of Technology, Surat, Gujarat, India

***Abstract*—In the current age, the attention of researchers is immersed by numerous imbalanced data applications. These application areas are intrusion detection in security, fraud recognition in finance, medical applications dealing with disease diagnosis pilfering in electricity, and many more. Imbalanced data applications are categorized into two types: binary and multiclass data imbalance. Unequal data distribution among data diverts classification performance metrics towards the majority data instance class and ignores the minority data, instance class. Data imbalance leads to an increase in the classification error rate. Random Forest Classification (RFC) is best suitable technique to deal with imbalanced datasets. This paper proposes the novel oversampling rate calculation algorithm as Improvised Dynamic Binary-Multiclass Imbalanced Oversampling Rate (IDBMORate). Experimentation analysis of the proposed novel approach IDBMORate on Page-block (Binary) dataset shows that instances of positive class is increased from 559 to 1118 whereas negative instance class remains same as 4913. In case of referred multiclass dataset (Ecoli), IDBMORate produces the consistent result as minority classes (om, omL, imS, imL) instances are oversampled majority class instances remains unchanged. IDBMORate algorithm reduces the ignorance of minority class and oversamples its data without disturbing the size of the majority instance class. Thus, it reduces the overall computation cost and leads towards the improvisation of classification performance.**

Keywords*—Binary imbalance; multiclass imbalance; oversampling; random forest classification; classification

I. Introduction

Numerous ranges of applications in today's real-world deal with imbalanced data applications. Numerous domains specifically medical diagnosis text mining, tracking of financial transactions, telecommunication, and industrial and engineering applications [1,2,3]. Dealing with these applications attracts researchers to resolve the data imbalance challenge. For the rapid development of real-world applications, information management with imbalanced classification is a decisive task. The upcoming needs of this digitized world comprise the utilization of technologies that can handle complex unevenness within the data sample distribution within data. There are a variety of functional application areas which need to reshape unbalanced, complex, and huge volumes of data by incorporating sampling techniques [4, 5, 6].

Data sampling methods are trendy in addressing class inequality at the data level and generally show improvement in classification results. The existing sampling approaches show that there is performance inconsistency if it is applied on both binaries as well as multiclass imbalance data application. The existing imbalanced data applications and work depict that there is an excessive sample generation in the existing oversampling methods which diverse the classification accuracy towards the majority data sample class [7,8]. It also increases the computation cost due to excessive sample generation. Present scenarios also have a diversion in data size of majority data sample in oversampling process and ignorance of minority data sample class. Data sample ignorance in the minority class leads to missing important information and overfitting in the majority class due to excessive data generation in the oversampling process. These challenges motivated this research work to derive a novel oversampling algorithm.

Imbalanced data classification biases performance towards majority numbered class in case of a binary application or majority classes in case of multiclass applications [9]. Traditional approaches lean towards abridged accuracy due to the massive amount of biased data towards the majority [10]. The proposed research work deals with a novel oversampling rate algorithm. In the existing study, the sampling methods which are suitable for the binary imbalance category are not suitable for multiclass imbalance application domains. The proposed IDBMORate algorithm is targeted to calculate oversampling rate which is dynamically applicable to binary as well as multiclass data imbalance and get enhanced classification performance.

In the first attempt, the proposed novel oversampling algorithm deals with the dynamicity of data oversampling which applies to both categories. The second advantage of the proposed algorithm is it will not disturb the majority data instance class and only focus on oversampling the minority data sample class. These two advantages indicate the strengths of the proposed algorithm in terms of less computation time and enhanced classification performance. The main objective of the paper is to identify imbalanced application areas and study existing sampling techniques. The subsequent objective of this research study is to propose a novel oversampling algorithm that leads to performance improvement. Experimental analysis of proposed IDBMORate on selected

binary and multiclass datasets shows improved performance metrics.

A. Organization of the Paper

The research study in this paper is organized as follows. The next section deals with a brief review of the related literature study of binary and multiclass imbalanced application domains and suitability of classifier. The third section emphasis on existing sampling approaches. Subsequent fourth section deals with the study of proposed Improvised Dynamic Binary-Multiclass Imbalanced Oversampling Rate (IDBMORate) algorithm and experimentation. Experiment analysis is carried on both binary (Page-block Dataset) as well as multiclass imbalanced (Ecoli Dataset) for verifying the dynamicity of proposed algorithm. Subsequent section deals with computational results of proposed IDBMORate. The final section outlines the major advantages and dynamicity of the proposed research work.

B. Research Gap

Excess time and computation cost required for generating new data samples for balancing the data. Proposed IDBMORate overcomes this research gap by oversampling minority class without disturbing majority data class and improvises classification performance.

II. Imbalanced Application Domains

This section of the paper focuses on imbalanced application domains and the suitability of the classifier for binary and multiclass imbalanced application domains [11,12]. It also highlights the issues raised due to data imbalance [13,14].

A. Imbalance Application and Suitability of Classifier

Classification with Imbalanced Dataset (ID) deals with heterogeneous and other imbalances with a massive amount of data.

Fig. 1 depicts the compatibility flow of classifiers depending upon the type of massive and streamed data. It shows that traditional classifiers are best suitable for balanced datasets [15,16] and Random Forest Classifier is best suitable for imbalanced data applications [24].

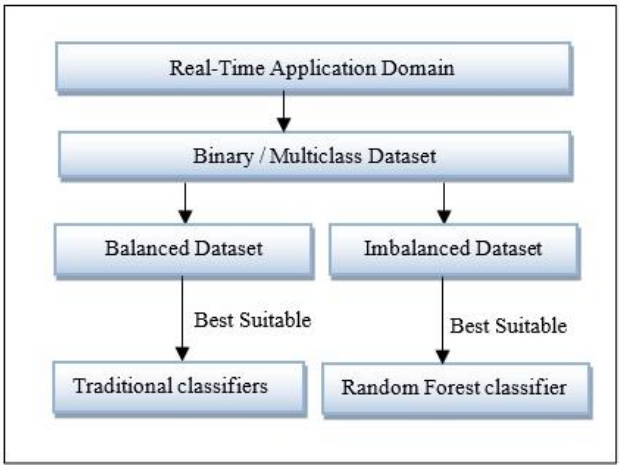


Fig. 1. Data Types and Suitability of Classifier.

B. Binary and Multiclass Imbalanced Application Domains

There is a list of numerous numbers of imbalanced applications which belong to class types as either binary imbalance or multiclass imbalance [17]. Table I nominates a list of selected applications with data domain analysis and categorization as binary, multiclass, or of both binary and multiclass imbalance. Binary classification techniques are the most progressive technique to deal with several applications such as medical diagnosis, and fault-finding activities in various business domains which always put forth the statistical results either belonging to one category of data or belonging to a second category [18,19,20].

TABLE I. Imbalance Applications with Domain and Category [11-22]

Sample Application		**Imbalance Category**
Diagnosis of cancer-infected patients and patient categorizing	Medical	Binary and Multiclass
Detection of an error occurring in code blocks in software projects	Software development	Binary class
Analyzing the count of faulty machines in industries	Industrial monitoring	Binary class
Multi-dimensional image categorization in various smart city applications	Hyperspectral image processing	Multiclass
Recognition of actions sequences and objects in videos	Mining of video	Binary and multiclass
Analyzing normal and dangerous actions	Action analysis	Binary class
Target specified classification with defined and varied frequency	Targeted classification domain	Multiclass
Analysis of literature relations in text	Mining of text	Binary class
Occurrence of frequent and rare activities in various domains	Activity recognition	Imbalance Multiclass
Recognition of annoyance and sentiment in text	Sentiment analysis	Binary and multiclass
Detection of normal and fraudulent transactions	Finance	Binary class
Categorization of deceptive and ordinary calls	Telecommunication	Binary class

To deal with the classification analysis of these binary and multiclass imbalance data applications, numerous approaches are discussed in the upcoming sections. Data imbalance approaches works at different data level or algorithmic level. At the data level based on the nature of the data, the approaches are categorized [23]. Table I summarizes selected applications, related application domains, and class categories.

III. Related Work

A. Existing Sampling Approaches

Sampling techniques are used to balance distorted data distribution Fig. 2 depicts categories of probability and non-probability sampling techniques [24].

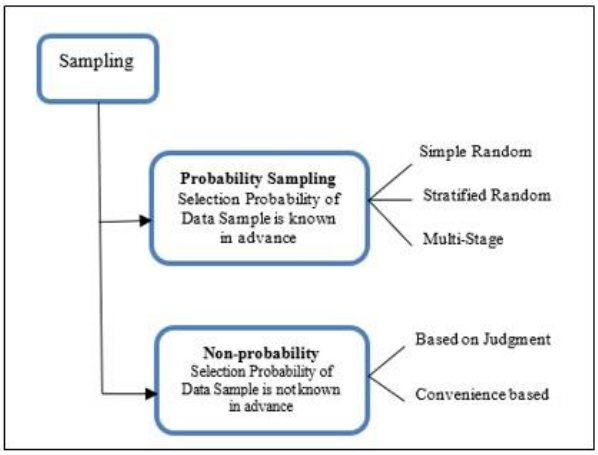


Fig. 2. Probability-based Sampling Strategies.

Both strategies have different sampling approaches to balance the dataset. Table II indicates the simple random sampling techniques steps [22], [23].

TABLE II. SIMPLE RANDOM SAMPLING ALGORITHMIC STEPS

Input: Imbalance data of sample size X provided	
Step:1	Take input as an imbalanced data set.
Steps:2	Distribution of dataset into x number of subsets with equal selection probability.

Table III indicates the stratified random sampling techniques steps [24].

TABLE III. STRATIFIED RANDOM SAMPLING ALGORITHMIC STEPS

Input: Imbalance data of sample size X provided	
Step:1	Take input as an imbalanced Data Set.
Step:2	Dataset distribution into "Strata".
Step:3	From each stratum select x as any random data sample.
Step:4	Merge the stratum x into the overall data sample.

Sample Case:

Game X has a team of 600 girl participants and 400 boy participant members. For applying a 30-number stratified random sample there is a need to select 12 boy participants from 400 and 18 girl participants from the overall count of 600 participants [25].

Table IV indicates the multistage sampling techniques steps.

TABLE IV. MULTI-STAGE SAMPLING ALGORITHMIC STEPS

Input: Imbalance data of sample size X provided	
Step:1	Take input as an imbalance data Set.
Step:2	Stage I sampling is based on one data attribute as selection criteria for all data samples provided in the data set.
Step:3	Stage II sampling is based on another data attribute as selection criteria for all data samples.

Sample case: Compilation of region-wise voters list based on numerous attributes like city, gender, etc. [26,27].

IV. PROPOSED ALGORITHM AND EXPERIMENTATION

This section of the paper deals with the evolution of the proposed algorithm Improvised Dynamic Binary-Multiclass Imbalanced Oversampling Rate (IDBMORate) to balance the imbalance ratio for both the category that is binary as well as multiple classes. The proposed algorithm targets the aim of oversampling minority data sample classes.

TABLE V. PROPOSED ALGORITHM

Algorithm: IDBMIORate	
	Input: Total # of Classes *C*, Distribution *D* Original Imbalanced Dataset *S*
	Output: Oversampling Rate of minority Class
Step 1	Calculate *n_min* through *D*
Step 2	*n_max = len(D)*
Step 3	Assign N= *S*
Step 4	*max = math. Ceil((n_max*(len(c)-1))/n_min)*
Step 5	Declare D, i=0, Declare N
Step 6	*while i < max*
Step 7	*Total samples = len(_N)*
Step 8	*Samples in min Sample Class=min(_D)* *pmin* =Calculate current ratio of minimum samples if (pmin < *(2/(3*lenI-1))) then* *current_min_class = sort_ values_D*
Step 9	End if Update Values of *_D and _N* *S[Class] = = current_min_class*
Step 10	._N = *append(samples_in_original_data)*
Step 11	_D = *sort_values(_N['Class']).*
Step 12	Update value of *i =i+1*
Step 13	Compute IOrate = _D – D
Step 14	*data = IOrate Data (IOrate, _N)*
Step 15	return *IOrate*

IDBMORate is successfully targeting data rescaling, selection of data, the invention of extra data, and transformation of data. The proposed algorithm deals with the dynamic approach of oversampling rate calculation.

Table V deals with the proposed IDBMIORate algorithm and Table VI deals with the related terminology.

Table VII deals with the data distribution of referred dataset.

TABLE VI. TERMINOLOGY USED FOR PROPOSED ALGORITHM

Key Term	**Specifications**
S	Original Imbalanced Dataset
C	Total number of Classes
N	Number of Total Data Sample Dataset
D	Data Distribution
n_{Min}	Number of Minority instance data Sample
n_{Max}	Count of Minority instance data Sample
RFC	Random Forest Classification

TABLE VII. DATA DISTRIBUTION SUMMARY

Dataset [28]	Total # of Instances	Imbalanced Category	# Data distribution according to classes
Page-blocks	5472	Binary	Positive 4913 Negative 559
Ecoli	336	Multiclass	cp 143, im 77, pp 52, imU 35, om 20, omL 5, imS 2, imL 2

A. Experimental Analysis of IDBMORate for Binary Datasets

For the Proposed Algorithm IDBMORate the experimentation has been carried out in Python Programming Platform for binary Dataset. Execution on Binary Imbalanced Dataset -1 is set Page-blocks with Random Forest Classification Model Total Data size: 5472.

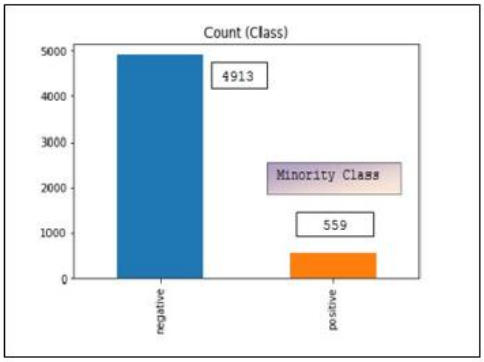


Fig. 3. Result before Sampling – Page-Block Dataset.

Fig. 3 specifies the distribution of class labels before sampling for class negative is 4913 and for class for the positive class are 559. Result after Sampling through IDBMIORate is as depicted in Fig. 4.

Fig. 5 deals with classification result of Page-block dataset with the proposed algorithm

Table VIII deals with the classification performance metrics of the proposed algorithm with RFC for the page block binary dataset.

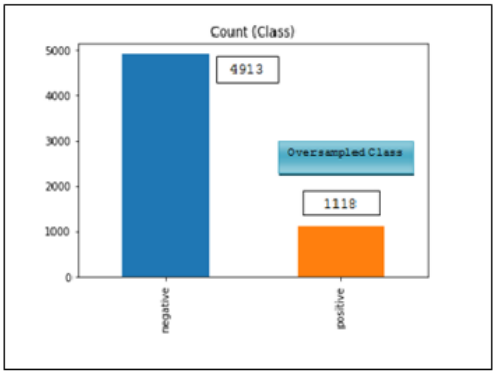


Fig. 4. Result with a Proposed Algorithm – Page-Block Dataset.

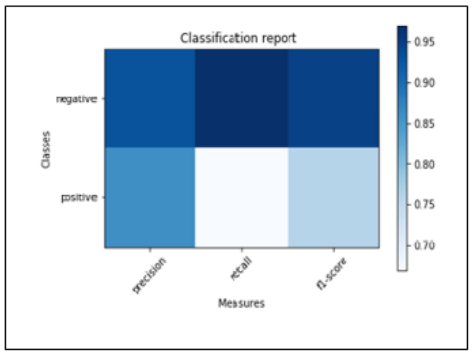


Fig. 5. Page-Block Dataset Classification Performance Graph.

TABLE VIII. PERFORMANCE METRICS FOR PAGE BLOCK DATASET

Performance Parameters / **Class**	**Precision**	**Recall**	**F1-Score**	**Support**
negative	0.93	0.97	0.95	1463
positive	0.86	0.67	0.76	347
Avg / Total	0.91	0.92	0.91	1810

B. Experimental Analysis of IDBMORate for Multiclass Datasets

The proposed algorithm also outperforms in the case of multiclass dataset. For performance evaluation of the multiclass dataset, this research study has used Ecoli dataset which contains multiple classes. The total sample size of the Ecoli dataset is 336.

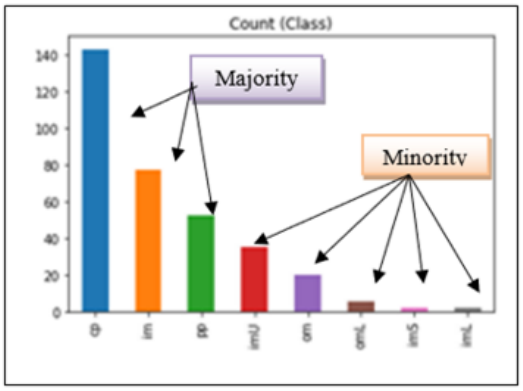


Fig. 6. Result before Sampling – Ecoli Dataset.

Fig. 6 indicates results before Sampling.

Distribution of class labels before Sampling for class cp 143

Distribution of class labels before Sampling for class im 77

Distribution of class labels before Sampling for class pp 52

Distribution of class labels before Sampling for class imU 35

Distribution of class labels before Sampling for class om 20

Distribution of class labels before Sampling for class omL 5

Distribution of class labels before Sampling for class imS 2

Distribution of class labels before Sampling for class imL 2

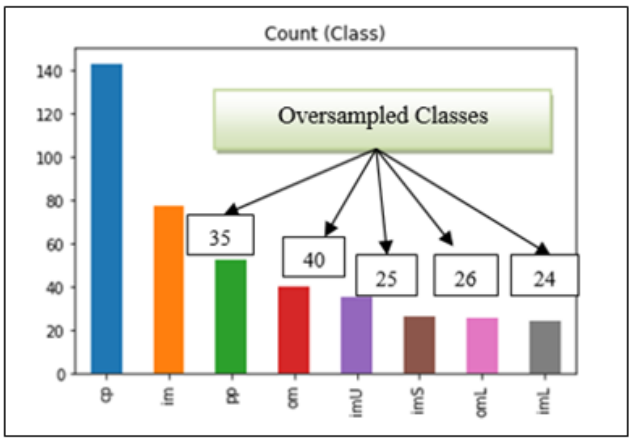


Fig. 7. Result with a Proposed Algorithm – Ecoli Dataset.

Fig. 7 indicates results after Sampling.

Distribution of class labels after Sampling for class cp 143

Distribution of class labels after Sampling for class im 77

Distribution of class labels after Sampling for class pp 52

Distribution of class labels after Sampling for class imU 35

Distribution of class labels after Sampling for class om 40

Distribution of class labels after Sampling for class omL 25

Distribution of class labels before Sampling for class imS 26

Distribution of class labels before Sampling for class imL 24

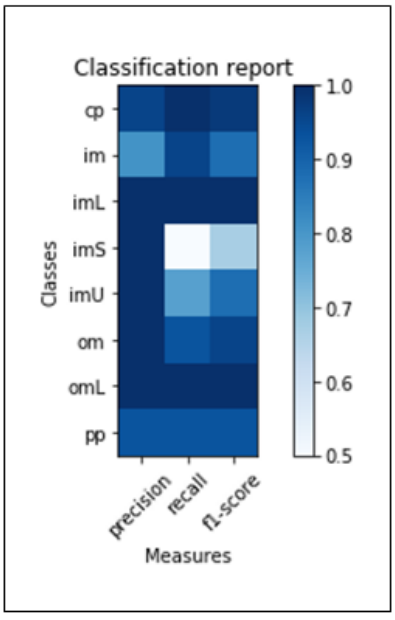


Fig. 8. Ecoli Dataset Classification Performance Graph.

Fig. 8 depicts the multiclass classification result with the proposed novel sampling approach. Precision, Recall, F1 Score, and support parameters are used for measuring the classification performance for both page block (binary) and Ecoli (Multiclass) dataset.

TABLE IX. PERFORMANCE METRICS WITH A PROPOSED ALGORITHM FOR MULTICLASS-ECOLI DATASET

Performance Parameters / Class	Precision	Recall	F1- Score	Support
cp	0.96	1.00	0.98	44
im	0.81	0.96	0.88	26
imL	1.00	1.00	1.00	4
imS	1.00	0.50	0.67	8
imU	1.00	0.78	0.88	9
om	1.00	0.93	0.96	14
omL	1.00	1.00	1.00	7

Table IX shows the RFC classification result with the proposed oversampling rate algorithm to compute the effectiveness of the proposed algorithm.

V. CONCLUSION AND FUTURE WORK

This research work addressed binary and multiclass imbalanced application domains, associated problems, and approaches to dissolve data imbalance dynamically. The proposed algorithm Improvised Dynamic Binary-Multiclass Imbalanced Oversampling Rate (IDBMORate) balances the minority classes without affecting the majority class which minimizes the cost of computation. Experimentation analysis on dataset page block and Ecoli has been carried out. IDBMORate algorithm overcomes the problem of the generation of extreme synthetic data samples for the minority classes, which leads to improved classification accuracy with the Random Forest Classification Model. Experimental analysis shows that IDBMORate efficiently outperforms the existing oversampling techniques for both binary as well as Multiclass imbalanced real-life scenarios. (IDBMORate) balances the minority classes without affecting the majority class which minimizes the cost of computation. The Proposed algorithms Improvised Dynamic Binary-Multiclass Imbalanced Oversampling Rate proposed algorithm which shows improvised results for both binary as well as multiclass DATA. THE hybrid sampling method will be focused in the future to upgrade the performance. The more dynamic method can be focused to work in a distributed environment.

REFERENCES

[1] Ahmed M. Sayed,"Machine Learning Augmented Breast Tumors Classification using Magnetic Resonance Imaging Histograms" International Journal of Advanced Computer Science and Applications, vol.12., no.12, pp.1-9, 2021.

[2] Prakruthi M K, Komarasamy G, "Novel Framework for Enhanced Learning-based Classification of Lesion in Diabetic Retinopathy", International Journal of Advanced Computer Science and Applications, vol.13., no.6, pp.37-44, 2022.

[3] Angelo, P., Resende, A.and Drummond, A. C. "A Survey of Random Forest Based Methods for Intrusion Detection Systems", ACM Comput. Surv. 51(3), 48-48.36,2018.

[4] Kamlesh Upadhyay, Prabhjot Kaur, Ritu Sachdeva," Fast and Robust Fuzzy-based Hybrid Data-level Method to Handle Class Imbalance", International Journal of Advanced Computer Science and Applications vol.13., no.6, pp.65-74, 2022.

[5] Delplace, A., Hermoso, S., and Anandita, K. ,"Cyber Attack Detection thanks to Machine Learning Algorithms", COMS7507: Advanced Security, 1-46, 2019.

[6] Jyoti Islam, Yanqing Zhang, "Brain MRI Analysis for Alzheimer's Disease Diagnosis Using an Ensemble System of Deep Convolutional Neural Networks", International Journal of Springer, 2018.

[7] Elyan, E., Francisco, C., Garcia, M. and Jayne, C, CDSMOTE: Class Decomposition and Synthetic Minority Class Oversampling Technique for Imbalanced Data Classification, Neural Computing & Applications, 33, 2839–2851,2020.

[8] Hamad, R. A., Kimura, M., and Lundström, J. Efficacy of Imbalanced Data Handling Methods on Deep Learning for Smart Homes Environments. SN COMPUT. SCI. 1, 204,2020.

[9] SMahmoud B. Rokaya, "Shallow Net for COVID-19 Classification based on Biomarkers", International Journal of Advanced Computer Science and Applications, vol.13., no.6, pp.97-103, 2022.

[10] Mohd Hakim Abdul Hamid, Marina Yusoff, Azlinah Mohamed "Survey on Highly Imbalanced Multi-class Data" International Journal of Advanced Computer Science and Applications, vol.13., no.16, pp.211-229, 2022.

[11] Evangeline D, Amy S Vadakkan, Sachin R S, Aakifha Khateeb, Bhaskar C5 "Cyberbullying Detection in Textual Modality" International Journal of Advanced Computer Science and Applications, vol.12., no.12, pp.217-229, 2021.

[12] Pramod Sunagar, Anita Kanavalli,"A Hybrid RNN based Deep Learning Approach for Text Classification", International Journal of Advanced Computer Science and Applications, vol.13., no.6, pp.289-295, 2022

[13] Ainul Yaqin, Majid Rahardi, Ferrian Fauzi Abdullah, "Accuracy Enhancement of Prediction Method using SMOTE for Early Prediction Student's Graduation in XYZ University," International Journal of Advanced Computer Science and Applications, vol.13. ,no.6, pp. 418-422, 2022.

[14] More, A. S., Rana, D. P., and Agarwal, IRandom Forest Classifier Approach for Imbalanced Big Data Classification for Smart City Application Domains. International Journal of Computational Intelligence & IoT, 1(2), 261-266,2018.

[15] Cao, L., and Shen, H. , Imbalanced Data Classification Using ImprovedClustering Algorithm and Under-sampling Method, In Proceedings of 20th International Conference on Parallel and Distributed Computing, Applications and Technologies, pp.361-366, 2019.

[16] Chee Keong Chan., Alexander W., "Development of a platform to explore network intrusion detection system for cyber security." Journal of Computer and Communications, Vol. 6, pp.1-11,2018.

[17] E.Abrahim., A.Saleem, T. Dao, Zhaoheng Liu, "Multiple-objective optimization and design of series-parallel systems using novel hybrid genetic algorithm meta-heuristic approach," World Journal of Engineering and Technology, Vol. 6, No.1 pp. 532-555,2018.

[18] Holewik, J., Schaefer, G., Korovin, I., "Imbalanced ensemble learning for enhanced pulsar identification," Proceedings of International Conference , pp.515-524, 2020.

[19] Jegierski H., Saganowski, S, "An outside the box'' solution for imbalanced data classification," IEEE Access, Vol. 8, pp. 125191-125209,2020.

[20] Khaja Mohammad Shahzad., Jong Sou Park, "Optimization of intrusion detection through fast hybrid feature selection," Proceedings of the Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies, pp.1-5, 2019.

[21] Kim, J., Jeong, J., and Shin, J, "M2m: Imbalanced classification via major-to-minor translation," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", pp.13893-13902, 2020.

[22] Q. Wang, "Imbalanced Classification Based on Over-sampling and Feature Selection," IEEE 5th International Conference on Cloud Computing and Big Data Analytics), pp. 325-330, 2020.

[23] Du, G.; Zhang, J.; Li, S.; Li, C. Learning from class-imbalance and heterogeneous data for 30-day hospital readmission. Neurocomputing 420, 27–35, 2021.

[24] Anjali S. More and Dipti P. Rana," Performance enrichment through parameter tuning of random forest classification for imbalanced data applications", Materials Today: Proceedings, Vol. 56, No. 6, pp. 3585-3593, 2022.

[25] Srinivasan, R.; Subalalitha, C. Sentimental analysis from imbalanced code-mixed data using machine learning approaches. Distrib. Parallel Databases, Vol. 39, pp.1–16, 2021.

[26] H.X. Guo, Y.J. Li, J. Shang, M.Y. Gu, and Y.Y. Huang, "Learning from class-imbalanced data: Review of methods and applications," Expert Syst. Appl., vol. 73, pp. 220–239, 2017.

[27] M. Bach, A.Werner, J. Zywiec, and W. Pluskiewicz, "The study of under- and oversampling methods' utility in the analysis of highly imbalanced data on osteoporosis," Inf. Sci., vol. 384, pp. 174–190, 2017.

[28] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.

Predicting Malicious Software in IoT Environment Based on Machine Learning and Data Mining Techniques

Abdulmohsen Alharbi, Md. Abdul Hamid, Husam Lahza
Department of Information Technology
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah 21589, Saudi Arabia

***Abstract*—The Internet of Things (IoT) enable the IoT to sense and respond using the power of computing to autonomously come up with the best solutions for any industry today. However, Internet of Things have vulnerabilities since it can be hacked by cybercriminals. The cybercriminals know where the IoT vulnerabilities are, such as unsecured update mechanisms and malware (Malicious Software) to attack the IoT devices. The recently posted IoT-23 dataset based on several IoT devices such as Philips Hue, Amazon Echo devices and Somfy door lock were used for machine learning classification algorithms and data mining techniques with training and testing for predictive modelling of a variety of malware attacks like Distributed Denial of Service (DDoS), Command and Control (C&C) and various IoT botnets like Mirai and Okiru. This paper aims to develop predictive modeling that will predict malicious software to protect IoT and reduce vulnerabilities by using machine learning and data mining techniques. We collected, analyzed and processed benign and several of malicious software in IoT network traffic. Malware prediction is crucial in maintaining IoT devices' safety and security from cybercriminals' activities. Furthermore, the Principal Component Analysis (PCA) method was applied to determine the important features of IoT-23. In addition, this study compared with previous studies that used the IoT-23 dataset in terms of accuracy rate and other metrics. Experiments show that Random Forest (RF) classifier achieved the predictive model produced classification accuracy 0.9714% as well as predict 8754 samples with various types of malware and obtained 0.9644% of Area Under Curve (AUC) which outperforms several bassline machine learning classification models.**

***Keywords*—Machine learning; internet of things; malware; predictive modeling; cyber threats**

I. INTRODUCTION

The Internet of Things are internet-connected devices that can transfer data over a network. Nowadays lots of cyber-attacks have increased, the cybercriminals seek to exploit or damage data, disrupt computer devices and network resources. The term cyber generally, defines computer devices, network, internet and information technology. [1] Cyber threat is a possibility to successful cyber-attack that aims to harm computer system or network, steal sensitive data and gain unauthorized access. However, the IoT are vulnerable in terms of security; cybercriminals use malware attacks such as DDoS, ransomware, and IoT botnet attack to disable systems and networks. Study by Gotsev et al. [2] used different machine learning models to evaluate the performance of Machine Learning (ML) for attack detection.

The researchers used all the features of IoT-23 dataset [3] which has 21 features and they detected different types of malware attack on IoT devices such as DDoS, Okiru, HorizontalPortScan and other IoT botnets. Similar study by Nicolas Stoian [4] focused on the security aspect of the IoT by investigating the usability of ML approaches on anomaly detection. In the research, the dataset has been split into 80% for training and 20% for testing for each ML algorithms. In results, the best ML algorithm is Random Forest with a weighted average precision of 100%.

Chunduri et al. [5] used multi class classification to detect IoT botnet malware. Their aim is to build a classifier to detect IoT botnet attack and to get the best accuracy possible by using machine learning classifiers. They have used Network Traffic Analysis Tool (Zeek) [6] that monitors all the traffic on network for malicious activity. The PCA method was applied to minimize features and maximize the accuracy rate by using ML classification algorithms such as Decision Tree (DT) K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Naïve Bayes (NB). The traditional approaches using static analysis method is complicated in terms of examining malicious software that exposes IoT to security breach risks. Therefore, it is important to improve the method by utilizing machine learning and data mining techniques to make it safer and more effective in predicting malicious software in IoT network traffic.

In this research the IoT-23 dataset [3] was used and collected, analyzed, processed to predict benign and several of malicious software in IoT network traffic. Furthermore, in results for prediction models, RF algorithm achieved highest accuracy for predicting malicious and benign in IoT network traffic. Section II is about background of cyber threats, the literature review is reviewed in Section III, Section IV is methodology, Section V describes performance evaluation of ML classification algorithms, Section VI describes experiments and results for validation of predictive models of malicious and benign in IoT network traffic and Section VII is about discussion of the results.

A. Problem Statement

Cybercriminals use malware attacks on IoT devices to hack and disable IoT devices. The attacks make IoT devices less efficient and cybercriminals can steal sensitive information and personal data. The cybercriminals know where the vulnerabilities of IoT devices are and exploit them through malware attacks on the devices. The existing studies focus on detecting threats and ignore the significance of predicting malware threats that increase IoT devices' vulnerabilities. Moreover, none of the studies included malicious software as a threat to security in IoT devices. The studies have focused on end-to-end devices or attack surfaces. This research aims to predict malicious software in IoT network traffic in order to protect IoT devices and reduce vulnerabilities by using ML and data mining techniques.

B. Importance of this Research

The harmful effects of malware into IoT environment are exposure of IoT to security breach risks and stealing of sensitive information and personal data. Therefore, we collected, analyzed and processed benign and several of malicious software in IoT network traffic of IoT-23 dataset [3]. Four types of malicious software were chosen out of IoT-23 and we considered those that have more significant effect on the IoT devices, which are DDoS, C&C and various botnets like Mirai and Okiru. Types of IoT devices in IoT-23 dataset are Echo device, Hue device and door lock device. It is extremely important to predict malware in IoT network traffic in order to protect IoT devices and reduce vulnerabilities by increasing the security level for IoT devices and to improve the environment and make it more motivational. This research will be using machine learning and data mining techniques to make it safer and more effective in predicting malicious software in IoT network traffic.

Our contributions in this work are:

- Minimize features and maximize the accuracy rate by using PCA method and ML classification algorithms.
- Propose prediction model to predict malicious and benign in IoT network traffic by using supervised learning.
- Increasing the security level for IoT to improve the environment and make it more motivational.

II. Background

This section provides related background information and context to explain the objectives and relevant field of research of this work. The concept, types, and IoT network activities of malicious software and botnet are discussed. Then, the concept of cybercriminals using malware attack is addressed. Finally, the most important malware attack that cause security risks on IoT environment is introduced.

A. Cyber Threats

Nowadays lot of cyber-attacks have increased, the cybercriminals seek to exploit or damage data, disrupt computer devices and network resources. The term cyber generally, defines computer devices, network, internet and information technology [7]. Cyber-attack attribution is technique that tracks, identifies, and lays blame on the criminal of a cyber-attack or other hacking exploit. Cyberspace is the environment of the internet that involving a global of computer network or the internet to enable communications and data exchange activities. Cyber Threat Intelligence (CTI) generally is relying on the collection of information and its analysis with current or potential attacks that is threatening policy of the organizations [8]. Cybercriminals can be internal or external to the organization that is facing cyberattack. An attack on the computer devices, network or system performed by person who has authorization access is known as an insider attack. An attack that originates exposures from outside the organization and attempt to exploit IT equipment are known as external attacks [1]. Cyber threat is possibility to successful cyber-attack that aims to harm computer system or network, steal sensitive data and gain unauthorized access. Some top cyber threats are illustrated as following [9] [10] [11] [12] [13]:

B. Malware (Malicious Software)

Malware is computer code designed to disrupt and disable such as stealing sensitive data or taking control of computer system. Malware (Malicious Software) has remained the most common cyber threats since 2014. Approximately four million samples of malware on different devices are detected by security organizations in 2017. The increase of malware samples have escalated malware attacks.

C. Ransomware Attack

Ransomware is a type of malware, which restricts access to user files or a computer system till the victim pays a ransom. Ransomware is significant cybersecurity threat since it uses techniques to avoid detection system to attack legitimate users. Ransomware can be considered a part of malware and it has been evaluated as a separate threat, although it belongs to the malware category. Moreover, Ransomware is considered the most significant cyber-attacks nowadays.

D. Distributed Denial of Service (DDoS) Attack

It is a cyber-attack which is an attempt to compromise the availability of computer devices or network resources to make them unavailable to the legitimate or normal users. DDoS attack is aimed to send massive amount of superfluous requests in order to deny the server from responding to the valid requests immediately. Denial of Service (DoS) attack can damage the target that rely on an online presence, while DDoS attack strikes a target with several resources and is harder to stop DDoS attack.

E. Cyber Espionage

It is type of cyber-attack which is an act of obtaining confidential information without permission from the user of the information for economic, political, military or personal objectives. It includes utilization of the internet or a computer network over utilized proxy server, malicious software including Trojan horse and spyware. The targets of this attack are government and commercial sectors. Cybercriminals develop new tools and techniques to increase the number of attacks and the degree of damage caused to its victims.

F. IoT Botnet Attack

The Internet of Things (IoT) bot is a variant of a traditional botnet that contains a group of compromised computers, smart devices and sensors connected to the internet. IoT botnet attack is used by cybercriminals for causing damage such as financial and for illegitimate purposes in terms of control of malicious actors. Over 41% of all attacks are due to the vulnerabilities of the IoT devices and IoT botnet attack contribute approximately 13% total of attacks in various other information technology industries.

III. Related Work

Study by Gotsev et al. [2] used different machine learning models to evaluate the performance of ML for attack detection. They applied various ML classifiers such as Support Vector Machine, Random Forrest Naïve Bayes, Logistic Regression and Decision Tree. In the experiments, the researchers used all feature of IoT-23 dataset [3] which has 21 features. Furthermore, IoT-23 dataset contains labeled information of benign and malicious IoT network traffic. They detected different types of malware attack on IoT devices such as DDoS, Okiru, HorizontalPortScan and other. In testing results, DT and RF achieved highest accuracy detection which was 1.00% and LR classifier achieved 0.76% accuracy, SVM achieved 0.74% accuracy, while NB classifier had unsatisfied result and achieved 58% accuracy.

Similar study by Nicolas Stoian [4] focused on the security aspect of Internet of Things networks by investigating the usability of ML approaches of anomaly detection. The researcher used 14 features of IoT-23 dataset and applied statistical correlation to dataset in order to eliminate the data which was irrelevant to the label column. Furthermore, the research splitting the dataset into 80% for training and 20% for testing for each ML algorithms. In results, the best ML algorithm is Random Forest with a weighted average precision of 100%, another algorithm is AdaBoost with precision of 86% while Support Vector Machine has precision of 60% and Naïve Bayes with a weighted average precision of 76% Chunduri et al. [5] used multi class classification to detect IoT botnet malware. Their aim is to build a classifier to detect IoT botnet attack and to get the best accuracy possible by using machine learning classifiers. The researcher used IoT-23 dataset [3] which contains benign and malicious network traffic of IoT devices. They focused on six types of botnet attack which are Mirai, Bashlite, Torii, Hakai, Okiru and Muhstik, moreover they used Zeek (Network Traffic Analysis Tool) [6] that monitors all the traffic on network for malicious activity. Furthermore, the researchers selected 12 features of IoT-23 dataset; in results they applied ML classifiers to training and testing IoT-23 dataset. The best accuracy was achieved by RF 99.88%, GradientBoosting produced 99.36% and K-Nearest Neighbors achieved 96.14% while Support Vector Machine with 94.72% can be considered the least fit model. In study by Strecker et al. [14], the researchers compared the effectiveness machine learning classifiers based cyber security techniques on the IoT-23 dataset. They used seven features of the IoT-23 dataset and applied RF, SVM and KNN algorithms for IoT cyber security in 2021. Their result for malware detection, the highest accuracy is of RF 92.27%, the second-best accuracy is KNN 89.80% and SVM achieved 83.52%. In 2018, Mirsky et al. [15] built an Intrusion Detection System (IDS) with autoencoders for detection of online anomaly called Kitsune. The researchers have developed attribute extractor that consists in the following attribute categories which are Socket, Network Jitter, Host-MAC&IP and Channel. They have demonstrated on their results anomaly detection of Mirai botnet malware on IoT devices. In 2018, Meidan et al. [16] introduced a dataset called N-BaIoT for Bashlite and Mirai botnet malware that considering as Kitsune [17] attributes which have been implemented on nine different IoT devices. Ferrag et al. [18] have investigated the way seven contemporary Artificial Neural Networks (ANN) approaches perform training of the CICIDS-2018 and the BoT-IoT datasets. They have provided the details on overall accuracy, training time by using Deep Learning (DL) detection rate.

Potluri et al. [19] evaluates a Convolutional Neural Network (CNN) based network intrusion detection techniques. They used the NSL-KDD and the UNSW-NB15 datasets. These datasets are converted into an image such as format as part of the process. The researchers build the three layers of CNN to label for the attacks. The study is compared the GoogLeNet and ResNet50 with designed CNN approach that achieved the satisfying results, with accuracy rate achieved 91.14% on the NSL-KDD dataset and 94.9% on the UNSW-NB15 dataset. De La Torre Parra et al. [20] proposed a method for detecting attacks at the back-side and client end at the same time. The client's site uses a CNN model with micro security for the detection of DDoS, botnets, and phishing attacks. The authors designed a joint training method for minimizing the resource utilization for detection of attacks in IoT devices and maximized the usability of extracted features for using the back-end server. The scope of the study is limited to using the CNN model for detecting URL-based attacks aimed at the client's IoT device and the RNN–LTSM model at the back-end server for the detection of malware attacks.

The focus of the study by Pastor et al. [21] is to provide measures for the detection of these malwares using passive network-based monitoring. Network flow features were identified for this purpose according to relevancy, and they were used with deep learning models and machine learning models. The researchers used some algorithms i.e. C4.5 Random Forest (RF) and Deep Neural Networks (DNN) to compare their performance. The main aim was to monitor crypto mining and the detection of real-time flow. This was done through testing these models in complex scenarios using real servers and connections that were encrypted. Various features were employed to demonstrate the efficiency of these models against crypto mining.

Study by Li et al. [22] focused on Command and Control (C2) server that is employed by using a Domain Generation Algorithm (DGA) in order to generate communication between C2 and malware. This cannot be easily countered by using traditional methods like blacklisting. The researchers provide the framework of machine learning in order to deal with these threats. Real-time data were collected for one year using real traffic, and a deep learning model was proposed for

the classification of domains of DGA. Results showed an accuracy of 95.89%, 97.79%, 92.45% and 95.21% for framework classification, DNN model, clustering at the second level and HMM prediction, respectively.

The study by Sarker [23] presented the Cyberlearning for binary classification model in order to detect anomalies and classification of multi-class model of cyber-attacks. Features that are correlated to this were selected for an analysis of comprehensive nature. The empirical data on the effectiveness of this model was analyzed. This model takes the binary classification into account for evaluating the effectiveness in detecting anomalies and other cyber-attacks. The techniques for machine learning were employed. For the hidden layers, a security model that is based on an artificial neural network was presented, and the effectiveness for these was evaluated using various techniques. Security datasets NSL-KDD and UNSW-NB15 were examined to employ an experimental analysis. The findings were believed to provide a good reference to future research in the same field.

Another study by Li et al. [24] used a detection system called Significant Permission IDentification (SigPID). This system is designed with three levels without extracting the usage of Android permissions. These levels include pruning permission data to identify malicious apps and then classifying those malwares using only 22 significant permissions. These permissions are then compared with the baseline approach, and the final outcomes indicates the precision of up to 90% in F-measure, accuracy and recall as well. Their dataset contains 2000 malware and the SigPID is determined to have an effectiveness of 93.62 in the detection of dataset malware and effectiveness of 91.4% in detecting unknown malware. The researcher by Karanja et al. [25] used a novel approach towards analyzing and classifying malware. This is done using texture features and classical classifiers of machine learning that apply to the IoT malware. A low computation approach was employed by converting the malware binaries into images. This broke the environmental dependencies and platform barriers, considering the analysis of images is not limited to platforms. A 95% and 88% accuracy were achieved through a K-nearest neighbor and random forest classifier. The results showed that this method is applicable for real-time settings and can be employed for flagging off known IoT malware using preprocessed features of the image in known malware. D. Li and Q. Li [26] used a mixture of attacks that use multiple generative methods and yield adversarial malware with multiple manipulation sets. The adversarial training is used with a manipulated set with large cardinality. The robustness of malicious software detection against twenty-six evasion attacks is based on five methods using gradient-based, gradient-free, obfuscation, a mixture of attacks, and transfer the attack. The proposed methods improved the performance, but more research is required in the area of adversarial malware detection.

A. Data Mining Techniques

Data mining approaches in Internet of Things (IoT) systems are integrated to discover in terms of a range of well-established knowledge patterns such as supervised, unsupervised, semi-supervised, and statistical approaches. These data mining approaches enable classification, prediction and regression of upcoming streaming data to be able to be visualizing the knowledge and activate the sensors and actuators of the IoT systems. Numerous crucial data mining techniques are illustrated as following [27] [28] [29] [30]:

B. Classification

Classification in data mining is a popular technique that splits data points into various classes which assigns items in a collection to target categories or classes. It allows to organize dataset of all types, including complicated and massive dataset as well as small and simple ones.

C. Regression

Regression is a type of data mining technique utilized to predict numeric values given in a particular part of dataset. The most popular types of regression are linear and logistic regressions algorithms of machine learning. Furthermore, other types of regression can be performed depending on their performance on an individual dataset.

D. Prediction

Prediction in data mining is to predict the unknown values or outcomes. Prediction techniques in data mining discovers the correlation among dependent and independent variables and the correlation between independent variables. It predicts the identity of single variable based on the current description of some other related variable.

IV. Methodology

In methodology section, the author will show the chosen techniques and tools to implement the approach for predicting malicious and benign IoT network traffic by using machine learning and data mining techniques. Furthermore, The IoT dataset-23 was selected which has a labeled malicious and benign on IoT network traffic. In addition, data preprocessing was applied and used Principal Component Analysis (PCA) method for feature selection to make it suitable for a machine learning model. To achieve the goals of building a usable supervised machine learning model for predicting malicious and benign IoT network traffic, this research will be applying IoT-23 dataset [3] targeting Weka tools [31] for ML model and data Orange tools [32] for data mining techniques.

A. IoT Dataset Selection

A large dataset IoT-23 published in January 2020 [3] has been identified. IoT-23 consists of a labeled dataset with malicious and benign IoT network traffic, types of IoT devices in IoT-23 dataset i.e. Echo device, Hue device and door lock device. The IoT-23 dataset created by the Avast AIC (Artificial Intelligence and Cybersecurity) laboratory which is help for researchers to develop machine learning algorithms. IoT-23 dataset has twenty malicious captures executed from different IoT devices, in which 11 malware labels and one benign label have existed in IoT network traffic. Four types of malicious software have been chosen of IoT-23 and we consider those that have more significant effect on the IoT devices, which are DDoS, C&C and various botnets like Mirai and Okiru. Fig. 1 shows distribution of main labeled after preprocessing on IoT-23 dataset.

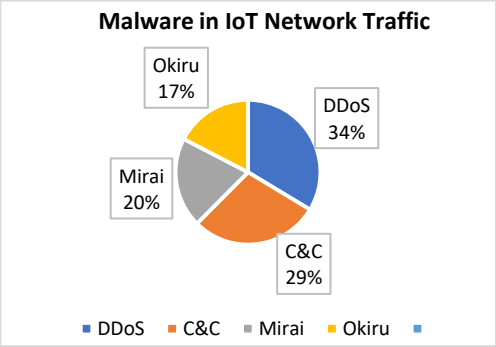


Fig. 1. Distribution of Main Labeled IoT-23 Dataset.

B. Feature Selection IoT-23 Dataset

After preprocessing IoT-23 dataset [3] the Principal Component Analysis (PCA) method was applied for feature selection. The IoT-23 dataset has 21 features and Weka tools [31] was used since it supports PCA method. After using PCA method, the IoT-23 dataset reduced to 18 features. The purpose of using PCA method is to find set of variables on IoT-23 dataset with less redundancy. Fig. 2 shows the main stages for feature selection using PCA method.

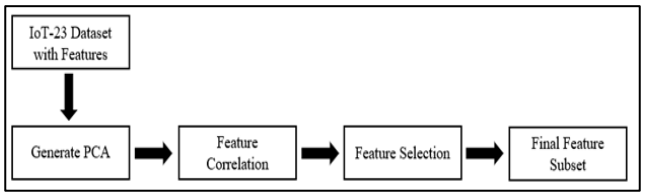


Fig. 2. Main Stages of the Feature Selection by PCA Method.

C. Principal Component Analysis

The PCA method was used for reducing the features of IoT-23 dataset. We eliminated these least important features for feature selection but do not lose original dataset completely. PCA method helps us to identify patterns in data of IoT-23 dataset based on the correlation among features. Furthermore, PCA method improved machine learning classification algorithms performance, removed correlated features and reduced overfitting by removing the unnecessary features in the IoT-23 dataset, which leads to minimizing features and maximizing the accuracy rate by using ML classification algorithms.

D. Supervised ML Models used for Prediction

In this research, the supervised learning model was used in terms of getting trained on a labelled dataset. A labelled IoT-23 dataset has two classes 0 and 1. 0 refers to benign while 1 refers to malicious, as binary classification we are predicting one of two classes in terms to know which features are malicious and benign of IoT network traffic. Fig. 3 shows prediction model for malicious and benign.

E. Weka Tools

Weka tools [31] is collection of machine learning algorithms for data mining tasks. It contains tools for data preprocessing, classification, clustering, regression and more. It is considered as an efficient tool for ML and data mining since it supports unsupervised and supervised ML algorithms.

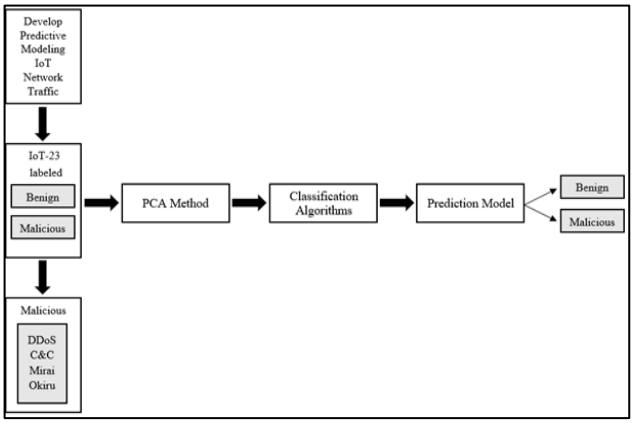


Fig. 3. Proposed Prediction Model.

F. Orange Tools

Orange tools [32] is an open-source platform to perform data analysis, machine learning and data mining Python scripting or visual programming. It contains prediction model that can predict the future based on previous attitude that happened before.

G. Classification Algorithms used in ML

Machine learning supervision was applied to train on a labelled IoT-23 dataset. A labelled dataset has two classes 0 and 1; 0 is benign and 1 is malicious. The labeled dataset is targeting to predict a packet which is malicious or benign. Furthermore, ML classification algorithms are applied since the labelled IoT-23 data has two classes. ML classification algorithms were used i.e. DT, RF, KNN, SVM and NB algorithms for prediction malicious and benign of IoT network traffic. Fig. 4 shows supervised ML classification algorithms used.

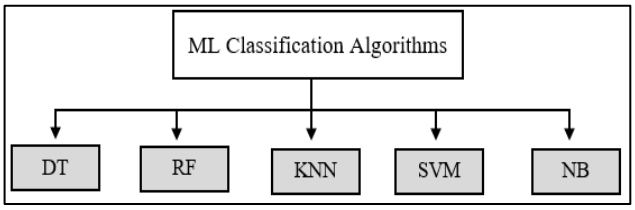


Fig. 4. Classification Algorithms used for Predication.

H. Training and Testing

For training and testing, the IoT-23 dataset was split into 70% for training and 30% for testing and 10-fold cross validation, to make it suitable for a machine learning models. Training or testing data is an approach to measure the accuracy of ML model. Moreover, machine learning classification algorithms were applied for training and testing model of ML.

V. Performance Evaluation

After training and validating ML models, some metrics are needed to identify the best model from a set of ML models. Our model of ML was developed to provide accurate prediction. Confusion matrix and evaluation metrics for classification model are used for predicting malicious and benign IoT network traffic. Additionally, four metrics used to

evaluate classification ML model which are accuracy, precision, recall and F1-score.

- **Accuracy** = $\frac{(TP+TN)}{(TP+FP+TN+FN)}$
- **Precision** = $\frac{TP}{(TP+FP)}$
- **Recall** = $\frac{TP}{(TP+FN)}$
- **F1-Score** = $\frac{2*(Precision*Recall)}{(Precision+Recall)}$

A. Confusion Matrix Model

Confusion Matrix Model (CMM) [33] [34] [35] [36] is applied to understand the performance of True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). Furthermore, CMM is used to describe the performance of the classifier model of test data for which the true values are known. Confusion matrix was applied for prediction malicious and benign as shown in Table I.

TABLE I. CONFUSION MATRIX

Confusion Matrix	Predicted Class: Benign	Predicted Class: Malicious
Actual Class: Benign	True Positive (TP)	False Negative (FN)
Actual Class: Malicious	False Positive (FP)	True Negative (TN)

B. Receiver Operating Characteristic (ROC) Curve

Receiver Operating Characteristic (ROC) curve was applied to show in a graphical way the trade-off between clinical sensitivity and specificity. The x-axis is false positive rate and the y-axis is true positive rate. Two metrics used to evaluate a ROC curve, Area Under the Curve (AUC) if equals 0.70% the model will be able to distinguish between true positive class and false positive class.

- **Sensitivity** = $\frac{TP}{(TP+FN)}$
- **Specificity** = $\frac{TN}{(TN+FP)}$

VI. EXPERIMENTS AND RESULTS

In the experiments, will present the results of the predicting malicious and benign of IoT network traffic by using various supervised machine learning classification algorithms. Four types of malicious software have predicted for each class which are DDoS, C&C, Mirai and Okiru. Performance model is evaluated using accuracy, precision, recall and F1-score. The IoT-23 dataset has split into 70% for training and 30% for testing and 10-fold cross validation, to make it suitable for a machine learning model. Experiments are done on prepared feature of IoT-23 dataset [3] using ML classification algorithms such as DT, RF, KNN, SVM and NB. This chapter also presents comparison of evaluation metrics for predictive model of the proposed technique with existing studies, for predicting malicious and benign IoT network traffic. The implementation of the ML model by Weka tools [31] and data mining techniques by Orange tools [32].

A. Malicious and Benign of IoT Network Traffic

The performance evaluation is computed using four metrics which are accuracy, precision, recall and F1-score. For training and testing by supervised learning (SL) the number of samples of malicious is 8K samples while benign is 43K samples as shown in Fig. 5 and Fig. 6. ML classification algorithms used i.e. DT, RF, KNN, SVM and NB for validation of malicious and benign in IoT network.

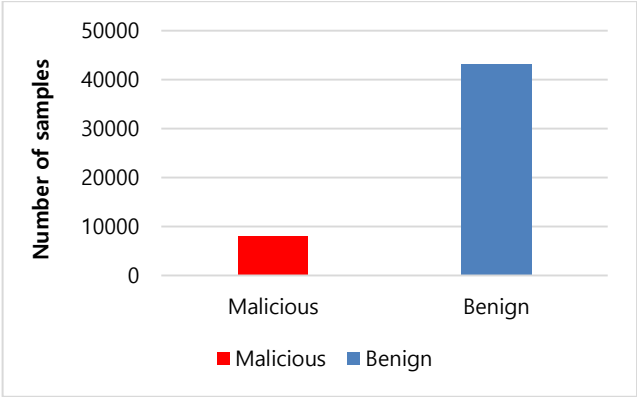


Fig. 5. Number of Samples Malicious and Benign.

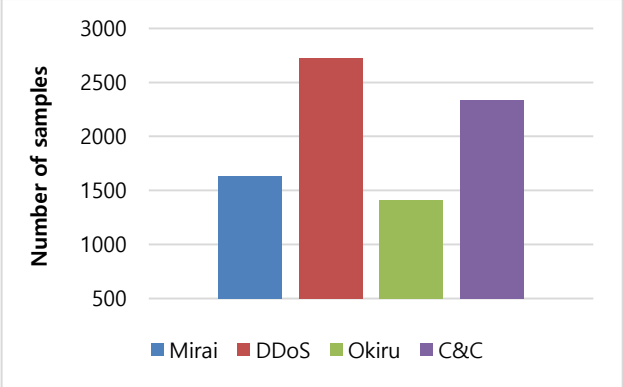


Fig. 6. Number of Samples each Type of Malware.

B. Performance Evaluation of ML Algorithms

Four metrics were applied for evaluating model of machine learning, the best results of ML classifiers algorithms were Random Forest and Support Vector Machine. Other ML algorithms obtained satisfying results, Table II shows performance evaluation of ML algorithms. Fig. 7 shows comparison performance metrics of ML algorithms, Fig. 8 comparison of ML algorithms using True Positive Rate (TPR).

TABLE II. PERFORMANCE EVALUATION OF ML ALGORITHMS

Classifier	Accuracy	Precision	Recall	F-1 Score
DT	0.9567	0.9580	0.9556	0.9553
RF	0.9848	0.9855	0.9850	0.9855
KNN	0.9674	0.9770	0.9773	0.9770
SVM	0.9840	0.9845	0.9850	0.9845
NB	0.9479	0.9668	0.9480	0.9544

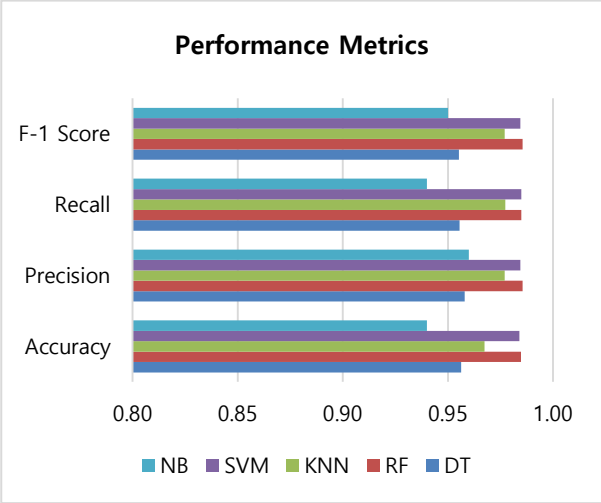


Fig. 7. Comparison Performance Metrics of ML Algorithms.

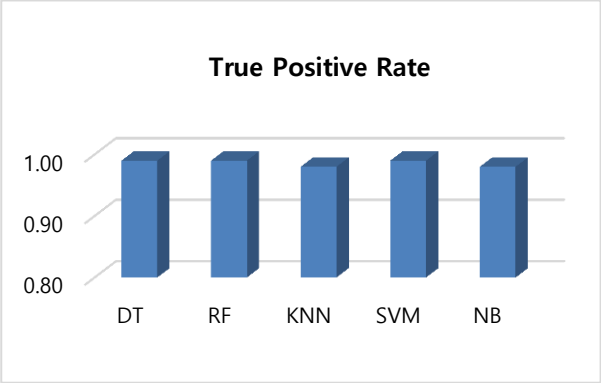


Fig. 8. Comparison of ML Algorithms using TPR.

C. Prediction Model for Malware of IoT Network Traffic

After training and testing ML model by Weka tools [31] Orange tools [32] was used for validating performance in terms of predicting malicious and benign in IoT network traffic. The results show RF algorithm is one the best accurate prediction mehods, this is due to the Classification Accuracy (CA) achieved 0.9714% while SVM algorithm obtained 0.7284% and we consider it obtained an inaccurate prediction, DT algorithm achieved 0.9141%, KNN obtained 0.9378% and NB obtained 0.8455%. As shown in Table III the number of samples for predicting malicious and benign in IoT network traffic. Fig. 9 shows a comparison of ML classifiers for predictive model accuracy.

TABLE III. VALIDATION A PREDICTION MODEL OF ML CLASSIFIERS

Classifier	Malicious	Benign
DT	12420	38854
RF	8754	42520
KNN	10733	40541
SVM	30753	20521
NB	21861	29413

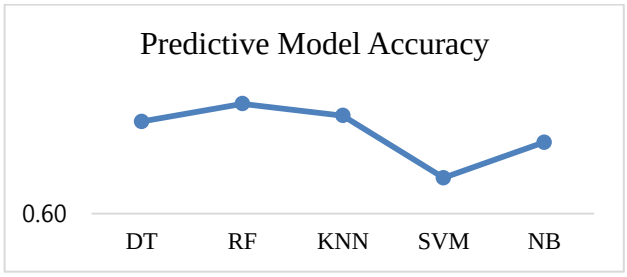


Fig. 9. ML Classifiers for Predictive Model Accuracy.

D. The Important Features for PCA Data Analysis

In Weka tools [31] Principal Component Analysis was used to reduce the features of the IoT-23 dataset [3]. We eliminate these least important features for feature selection but we do not lose original dataset completely. PCA method helps us to identify patterns in data of IoT-23 dataset based on the correlation among features. Furthermore, PCA method improved machine learning classification algorithms performance, removed correlated features and reduced overfitting by removing the unnecessary features in the IoT-23 dataset, which leads to minimizing features and maximizing the accuracy rate by using ML classification algorithms. PCA method considered the important features of IoT-23 dataset which are Ts, Uid, ID_orig.h, ID_orig.p, ID_resp.h, ID_resp.p, Proto, Service, Duration, Resp_bytes, Conn_state, Local_orig, Local_resp, Missed_bytes, History, Orig_pkts, Resp_pkts and Tunnel_parents. Fig. 10 shows a scatter plot after applied PCA method on IoT-23 dataset.

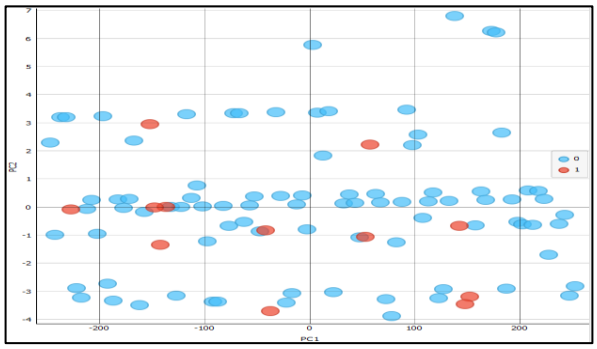


Fig. 10. Scatter Plot of Label IoT-23 Dataset.

E. Comparison with Previous Studies for Evaluating ML Models

The comparison of the machine learning algorithms with existing behavioural-based IoT-23 dataset is showed in Table IV. Gotsev et al. [2] used DT and RF classifiers and achieved 1.00% with four metrics. Nicolas Stoian [4] employed AdaBoost algorithm and it did not perform well as it achieved only 0.87% accuracy. Chunduri et al. [5] used RF and GBM algorithms and obtained highest accuracy rate, RF produced 0.9988% and GBM produced 0.9936%. Strecker et al. [14] obtained of RF classifier 0.9227% accuracy. Our model of ML produced the accuracy 0.9848% using RF classifier. In addition, model of ML achieved satisfied results for all ML classification algorithms with all metrics i.e. AUC, accuracy, precision, recall and f-1 score.

F. Confusion Matrix

After applying different ML classifiers, confusion matrix was applied i.e. True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). The x-axis describes predicted label and the y-axis describes true label. The results show as per DT classifier, 12999 predicted a packet is benign and it actually is, 1712 predicted a packet is malicious and it actually is not, 667 predicted a packet is benign but it actually is not, 4 predicted a packet is malicious but it actually is, as well as the other ML classifiers i.e. RF, KNN, SVM and NB shown in Table V.

TABLE IV. COMPARISON WITH PREVIOUS STUDIES FOR EVALUATING ML MODEL

Study	Model	AUC	Accuracy	Precision	Recall	F-1 Score
[2] (Gotsev et al. 2021)	Naïve Bayes	-	0.58	0.75	0.58	0.51
	Support Vector Machine	-	0.74	0.70	0.74	0.70
	Logistic Regression	-	0.76	0.74	0.76	0.73
	Decision Tree	-	1.00	1.00	1.00	1.00
	Random Forest	-	1.00	1.00	1.00	1.00
[4] (Nicolas Stoian 2020)	Support Vector Machine	-	0.67	0.60	0.67	0.59
	Naïve Bayes	-	0.23	0.27	0.38	0.10
	Artificial Neural Network	-	0.66	0.71	0.66	0.52
	Random Forest	-	0.84	0.88	0.85	0.84
	Adaptive Boosting	-	0.87	0.86	0.87	0.83
[5] (Chunduri et al. 2021)	K-Nearest Neighbors	0. 9568	0.9614	-	-	-
	Random Forest	0.9960	0.9988	-		
	Support Vector Machine	0.9400	0.9472	-	-	-
	Gradient Boosting Machine	0.9867	0.9936	-	-	-
[14] (Strecker et al. 2021)	K-Nearest Neighbors	0.8982	0.8990	0.8982	0.8971	0.9280
	Random Forest	0.9193	0.9227	0.9193	0. 9330	0. 9393
	Support Vector Machine	0.8352	0.8352	0.8352	0.8298	0.8559
Our Study	Decision Tree	0.9422	0.9567	0.9580	0.9556	0.9553
	Random Forest	0.9644	0.9848	0.9855	0.9850	0.9855
	K-Nearest Neighbors	0.9583	0.9674	0.9770	0.9773	0.9770
	Support Vector Machine	0.9628	0.9840	0.9845	0.9850	0.9845
	Naïve Bayes	0.9161	0.9479	0.9668	0.9480	0.9544

TABLE V. CONFUSION MATRIX FOR EACH ML CLASSIFIERS

DT Classifier		
True Label	Benign	Malicious
Benign	12999	4
Malicious	667	1712
	Predicted Label	

RF Classifier		
True Label	Benign	Malicious
Benign	12950	53
Malicious	180	2199
	Predicted Label	

KNN Classifier		
True Label	Benign	Malicious
Benign	12748	255
Malicious	246	2133
	Predicted Label	

SVM Classifier		
True Label	Benign	Malicious
Benign	12890	61
Malicious	210	2221
	Predicted Label	

NB Classifier		
True Label	Benign	Malicious
Benign	12871	155
Malicious	239	2117
	Predicted Label	

VII. Discussion of the Results

The IoT-23 dataset has approximately 160K rows and 21 features from 20 malware traffic captured from different IoT devices i.e. Echo device, Hue device and door lock device in which 11 malware labels and one benign label have existed in IoT network traffic. This study found the RF classifier to be the best performing; produced an AC 0.9714% and AUC achieved 0.9644%. For predicting malicious software over the IoT network traffic, all ML algorithms were predicting well except SVM algorithm this is due to AC produced was 0.7284%. DT algorithm predicted 12420 of malware which predict DDoS C&C, Mirai and Okiru; as well as the other ML algorithms have predicted malware i.e. RF, KNN, SVM and NB. Moreover, PCA method helps to improve ML performance and decrease overfitting by removing the unnecessary features in the IoT-23 dataset in order to improve accuracy rate for prediction. The IoT-23 dataset was split into 70% for training and 30% for testing and 10-fold cross validation, to make it suitable for a machine learning model. This study had two limitations. First, the types of malicious software in the IoT-23 dataset is limited. However, as discussed previously, four types of malicious software have been chosen on the IoT-23 dataset. However, we consider these types of malicious software selected have more significant effect in IoT environment. Second, after applied predictive modelling, malicious software cannot be prevented on the IoT devices. This is because the experiments were performed by machine learning and data mining techniques for predictive modeling without preventing tools for malicious software such as Intrusion Detection System (IDS).

VIII. Conclusion and Future Work

In this research, ML classification algorithms and data mining techniques were used for predictive modeling for validation of prediction of malicious and benign in IoT network traffic. Types of malware and IoT botnet used in this study for predicting are DDoS, C&C and various IoT botnet like Mirai and Okiru. The PCA method was applied to determine the important features of IoT-23 dataset and this study has been compared with previous studies that used the IoT-23 dataset in terms of accuracy rate and other metrics. We achieved better accuracy rate of ML classification i.e. KNN, SVM and NB. The highest accuracy rate for models of ML is RF classifier which produced 0.9844% and SVM classifier produced 0.9840%. For prediction model of malicious and benign in IoT network traffic, RF algorithm obtained the best accurate predictive model and achieved AC 0.9714% and predicted 8754 samples of various types of malware such as DDoS, C&C and various IoT botnet like Mirai and Okiru. In future work, the researchers will extensively understand the behavior of various types of malware attacks in IoT. Furthermore, will study these types of malware attacks against machine learning algorithms-based IDS. Also will investigate and evaluate IDS to prevent malicious software in network traffic.

References

[1] R. Saxena and E. Gayathri, "Cyber threat intelligence challenges: Leveraging blockchain intelligence with possible solution," Mater. Today Proc., p. S2214785321045752, Jul. 2021, doi: 10.1016/j.matpr.2021.06.204.

[2] L. Gotsev, M. Dimitrova, B. Jekov, E. Kovatcheva, and E. Shoikova, "A Cybersecurity Data Science Demonstrator: Machine Learning in IoT Network Security," p. 6, 2021.

[3] "Sebastian Garcia, Agustin Parmisano, & Maria Jose Erquiaga. (2020). IoT-23: A labeled dataset with malicious and benign IoT network traffic. doi: 10.5281/zenodo.4743746.

[4] N.-A. Stoian, "Machine Learning for Anomaly Detection in IoT networks: Malware analysis on the IoT-23 Data set," p. 10.

[5] H. Chunduri, T. Gireesh Kumar, and P. V. S. Charan, "A Multi Class Classification for Detection of IoT Botnet Malware," in Computing Science, Communication and Security, vol. 1416, N. Chaubey, S. Parikh, and K. Amin, Eds. Cham: Springer International Publishing, 2021, pp. 17–29. doi: 10.1007/978-3-030-76776-1_2.

[6] Zeek Network Security Monitor (2019). https://docs.zeek.org/en/current/intro/.

[7] M. S. Abdullah, A. Zainal, M. A. Maarof, and M. Nizam Kassim, "Cyber-Attack Features for Detecting Cyber Threat Incidents from Online News," in 2018 Cyber Resilience Conference (CRC), Putrajaya, Malaysia, Nov. 2018, pp. 1–4. doi: 10.1109/CR.2018.8626866.

[8] U. Noor, Z. Anwar, T. Amjad, and K.-K. R. Choo, "A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise," Future Gener. Comput. Syst., vol. 96, pp. 227–242, Jul. 2019, doi: 10.1016/j.future.2019.02.013.

[9] K. Alieyan, M. M. Kadhum, M. Anbar, S. U. Rehman, and N. K. A. Alajmi, "An overview of DDoS attacks based on DNS," in 2016 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Oct. 2016, pp. 276–280. doi: 10.1109/ICTC.2016.7763485.

[10] A. Alzahrani, A. Alshehri, R. Alharthi, H. Alshahrani, and H. Fu, "An Overview of Ransomware in the Windows Platform," in 2017 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, Dec. 2017, pp. 612–617. doi: 10.1109/CSCI.2017.106.

[11] M. Libicki, "The coming of cyber espionage norms," in 2017 9th International Conference on Cyber Conflict (CyCon), Tallinn, May 2017, pp. 1–17. doi: 10.23919/CYCON.2017.8240325.

[12] L. L. Dhirani, E. Armstrong, and T. Newe, "Industrial IoT, Cyber Threats, and Standards Landscape: Evaluation and Roadmap," Sensors, vol. 21, no. 11, p. 3901, Jun. 2021, doi: 10.3390/s21113901.

[13] A. P. Namanya, A. Cullen, I. U. Awan, and J. P. Disso, "The World of Malware: An Overview," in 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud), Barcelona, Spain, Aug. 2018, pp. 420–427. doi: 10.1109/FiCloud.2018.00067.

[14] S. Strecker, R. Dave, N. Siddiqui, and N. Seliya, "A Modern Analysis of Aging Machine Learning Based IoT Cybersecurity Methods," J. Comput. Sci. Appl., vol. 9, no. 1, pp. 16–22, Oct. 2021, doi: 10.12691/jcsa-9-1-2.

[15] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection," ArXiv180209089 Cs, May 2018, Accessed: Mar. 02, 2022. [Online]. Available: http://arxiv.org/abs/1802.09089.

[16] Y. Meidan et al., "N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders," IEEE Pervasive Comput., vol. 17, no. 3, pp. 12–22, Jul. 2018, doi: 10.1109/MPRV.2018.03367731.

[17] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," J. Inf. Secur. Appl., vol. 50, p. 102419, Feb. 2020, doi: 10.1016/j.jisa.2019.102419.

[18] B. Yan and G. Han, "Effective Feature Extraction via Stacked Sparse Autoencoder to Improve Intrusion Detection System," IEEE Access, vol. 6, pp. 41238–41248, 2018, doi: 10.1109/ACCESS.2018.2858277.

[19] S. Potluri, S. Ahmed, and C. Diedrich, "Convolutional Neural Networks for Multi-class Intrusion Detection System," in Mining Intelligence and Knowledge Exploration, vol. 11308, A. Groza and R. Prasath, Eds. Cham: Springer International Publishing, 2018, pp. 225–238. doi: 10.1007/978-3-030-05918-7_20.

[20] G. De La Torre Parra, P. Rad, K.-K. R. Choo, and N. Beebe, "Detecting Internet of Things attacks using distributed deep learning," J. Netw.

Comput. Appl., vol. 163, p. 102662, Aug. 2020, doi: 10.1016/j.jnca.2020.102662.

[21] A. Pastor et al., "Detection of Encrypted Cryptomining Malware Connections With Machine and Deep Learning," IEEE Access, vol. 8, pp. 158036–158055, 2020, doi: 10.1109/ACCESS.2020.3019658.

[22] Y. Li, K. Xiong, T. Chin, and C. Hu, "A Machine Learning Framework for Domain Generation Algorithm-Based Malware Detection," *IEEE Access*, vol. 7, pp. 32765–32782, 2019, doi: 10.1109/ACCESS. 2019.2891588.

[23] I. H. Sarker, "CyberLearning: Effectiveness analysis of machine learning security modeling to detect cyber-anomalies and multi-attacks," Internet Things, vol. 14, p. 100393, Jun. 2021, doi: 10.1016/j.iot.2021 .100393.

[24] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-an, and H. Ye, "Significant Permission Identification for Machine-Learning-Based Android Malware Detection," IEEE Trans. Ind. Inform., vol. 14, no. 7, pp. 3216–3225, Jul. 2018, doi: 10.1109/TII.2017.2789219.

[25] E. M. Karanja, S. Masupe, and M. G. Jeffrey, "Analysis of internet of things malware using image texture features and machine learning techniques," Internet Things, vol. 9, p. 100153, Mar. 2020, doi: 10.1016/j.iot.2019.100153.

[26] D. Li and Q. Li, "Adversarial Deep Ensemble: Evasion Attacks and Defenses for Malware Detection," IEEE Trans. Inf. Forensics Secur., vol. 15, pp. 3886–3900, 2020, doi: 10.1109/TIFS.2020.3003571.

[27] M. M. Gaber et al., "Internet of Things and data mining: From applications to techniques and systems," WIREs Data Min. Knowl. Discov., vol. 9, no. 3, May 2019, doi: 10.1002/widm.1292.

[28] H. A. Madni, Z. Anwar, and M. A. Shah, "Data mining techniques and applications — A decade review," in 2017 23rd International Conference on Automation and Computing (ICAC), Huddersfield, United Kingdom, Sep. 2017, pp. 1–7. doi: 10.23919/IConAC .2017.8082090.

[29] F. Ali, D. Bhatt, T. Choudhury, and A. Thakral, "A Brief Analysis of Data Mining Techniques," in 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, Dec. 2019, pp. 752–758. doi: 10.1109/ICCIKE47802.2019.9004252.

[30] I. Batra, S. Verma, and K. Janjua, "Performance Analysis of Data Mining Techniques in IoT," in 2018 4th International Conference on Computing Sciences (ICCS), Jalandhar, Aug. 2018, pp. 194–199. doi: 10.1109/ICCS.2018.00039.

[31] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[32] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, Journal of Machine Learning Research 14(Aug): 2349−2353.

[33] P. Bedi et al., "Detection of attacks in IoT sensors networks using machine learning algorithm," Microprocess. Microsyst., vol. 82, p. 103814, Apr. 2021, doi: 10.1016/j.micpro.2020.103814.

[34] A. Sivanathan, H. Habibi Gharakheili, and V. Sivaraman, "Managing IoT Cyber-Security Using Programmable Telemetry and Machine Learning," IEEE Trans. Netw. Serv. Manag., vol. 17, no. 1, pp. 60–74, Mar. 2020, doi: 10.1109/TNSM.2020.2971213.

[35] H. Naeem et al., "Malware detection in industrial internet of things based on hybrid image visualization and deep learning model," Ad Hoc Netw., vol. 105, p. 102154, Aug. 2020, doi: 10.1016/j.adhoc. 2020.102154.

[36] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Robust Intelligent Malware Detection Using Deep Learning," IEEE Access, vol. 7, pp. 46717–46738, 2019, doi: 10.1109/ACCESS.2019.2906934.

Power user Data Feature Matching Verification Model based on TSVM Semi-supervised Learning Algorithm

Yakui Zhu*, Rui Zhang, Xiaoxiao Lu
Marketing Service Center, State Grid Hebei Electric Power Co. Ltd, Shijiazhuang, China

***Abstract*—The existing model for identifying user data features based on smart meter data adopts a supervised learning method. Although the model has good identification performance under the condition of sufficient index samples, matching data are difficult to obtain and the marking cost is high in real life. The identification accuracy is significantly reduced when the matching data are insufficient or unavailable in the supervised learning method. In view of the above problems, based on the smart meter data, this paper proposes a feature recognition method for residential user data based on semi-supervised learning, which uses three indicators to evaluate the recognition performance of the proposed semi-supervised learning method for residential user data features and to find the appropriate feature selection method and data acquisition resolution. Then, explore the role of this method in real life when there is insufficient or unavailable matching data. Experimental results show that the performance of the proposed semi-supervised learning algorithm is better than that of the supervised learning algorithm, and the accuracy of the proposed algorithm is better than or close to that of the supervised learning algorithm.**

***Keywords*—Power system; data matching; data characteristics; semi-supervised learning algorithm; load model**

I. INTRODUCTION

With the continuous development of smart electricity business such as demand response and energy efficiency management, scholars at home and abroad have conducted in-depth exploration and research on the relationship between smart meter data and residential user data characteristics. The characteristic information of residential user data affects the electricity consumption behavior of residents, and conversely, the portrait information of residents can also be identified from their electricity consumption behavior and data [1]. Therefore, the topic of exploring the potential correlation between smart meters and residential user data characteristics is mainly divided into two categories: one is to explore the residential electricity load pattern and analyze how the user data characteristics affect the electricity load type; the other is to identify the residential user data characteristics through the residential electricity load characteristics [2].

The residential load has strong randomness, and the characteristic data of residential users can help power companies to better understand the characteristics of residential peak load and the reasons for changes in electricity consumption behavior. Scholars in China and abroad have done a lot of research on the first type of topic [3]. Wang Yi et al. proposed a new dynamic clustering method for power consumption behavior. Firstly, the symbolic aggregation approximation measure is used for each user to reduce the size of the data set, and the Markov algorithm based on time series is used to establish a dynamic energy consumption model, which converts a large number of load data curves into matrix form. Secondly, the typical power consumption pattern is obtained through the fast clustering algorithm based on the density peak, and then the Kullback Liebler distance is used to evaluate the difference in power consumption behavior, and the users are clustered. The example in this paper verifies the effectiveness of this method [4]. Wang Fei et al. used the density-based clustering algorithm DBSCAN to extract the seasonal typical electricity consumption patterns of each user and used the K-means clustering algorithm to cluster the electricity consumption patterns, and finally used the association rule mining algorithm to explore the potential relationship between the residential electricity consumption pattern and its user data characteristic factors [5]. These works deeply explore the factors affecting the electricity load pattern of users, which can effectively promote the implementation of energy-saving projects. However, these studies need to integrate a large number of user data characteristics, classify and manage users according to specific user data characteristics rather than load patterns, and make some personalized service policies for different types of families, which are more easily accepted and understood by non-professional technicians [6]. Therefore, how to automatically, intelligently, and accurately identify the characteristic information of resident user data has become the core work of the association. In recent years, there are many methods applied to identify the characteristics of residential user data in smart meter data.

In the research of user portrait recognition based on the smart meter data, the key task is to extract and select the features of smart meter data, which directly affects the upper limit of the performance of the user portrait recognition model [7]. In the existing literature, the extraction of smart meter data features is focused on a single domain (single time domain or single frequency domain), which fails to comprehensively analyze the potential rules contained in smart meter data from multiple perspectives [8].

However, the existing models for identifying user data features based on smart meter data all adopt the supervised learning method. Although it achieves good recognition performance when the index sample is sufficient, the recognition accuracy is significantly reduced when the

*Corresponding Author.

matching data is insufficient or unavailable [9]. However, in real life, it is difficult to obtain matching data, the cost is high, and it is time-consuming and labor-intensive. How to save the cost of sample labeling while maintaining good user data feature recognition performance is an urgent problem to be solved [10].

To solve the problems in real life, such as difficulty in obtaining matching data, high cost, time-consuming, and labor-consuming, this paper proposes a feature recognition method of residential user data based on semi-supervised learning based on extracting the time domain and frequency domain features of smart meters. This method can make full use of the potential rules contained in a small number of matching data and a large number of non-index data to explore the relationship between smart meter data and residential user data characteristics and reduce the cost of index marking. In this paper, the effectiveness of semi-supervised learning is verified by the real CER data set, and two main factors affecting the performance of the semi-supervised learning recognition algorithm are analyzed.

The main innovations of this paper are:

1) It decomposes an original average daily power consumption curve by adopting a discrete wavelet transform and extracting frequency domain characteristics;

2) The method of combining time domain and frequency domain features is helpful to improve the accuracy of portrait recognition.

3) The applicability and expansibility of the feature extraction method. Based on the combination of time domain and frequency domain characteristics, the resolution of smart meter data acquisition will have an impact on the results of portrait recognition.

The main contents of the paper are as follows:

1) It introduces the research motivation, background and research status of this paper, and puts forward the solutions to the existing problems.

2) It explains the basic principle of the technical content.

3) The data feature recognition method based on semi-supervised learning is introduced.

4) Through the experimental analysis, the technical advancement and reliability of the research content of this paper are compared and tested.

5) In the conclusion part, the research results of this paper and the future work are summarized.

II. Basic Principles of Semi-supervised Classification

Semi-supervised generative classification algorithms assume that different classes of data are generated by potentially different "sources", and that the samples of each class follow a probability distribution $p(x \mid y;\theta)$, where θ is the parameter of the probability density distribution function [11]. If the samples without index and the samples with index come from the same probability distribution, the samples without index whose index values are inferred can be used as training samples to improve the classification accuracy of the model [12].

The semi-supervised expectation-maximization (EM) algorithm is a typical generative classification model, which assumes that its data distribution conforms to the Gaussian mixture model, and that the data of each class follows the normal distribution [13]. The joint probability density of the sample data and the index can be obtained from the conditional probability density function of the class, as shown in (1).

$$p(x, y \mid \theta) = p(y \mid \theta)\, p(x \mid y, \theta) \tag{1}$$

The parameter vector of the model can be determined according to the samples with and without indicators, which is transformed into solving the optimization problem as shown in (2) [14].

$$\max_{\theta}(\ln p(\{x_i, y_i\}_{i=1}^{l} \mid \theta) + \lambda \ln p(\{x_i, y_i\}_{i=l+1}^{l+u} \mid \theta)) \tag{2}$$

Where, λ is the artificially set parameter. $x_1, x_2, \cdots, x_l$ is the sample with index, and $x_{l+1}, x_{l+2}, \cdots, x_{l+u}$ is the sample without index. The EM algorithm is used to solve this problem. First, a set of parameters is estimated according to the sample with index, and then the expectation is calculated (step E) [15]. The sample without index is marked according to the estimated parameters, and the maximum likelihood estimate is calculated. Maximize the maximum likelihood estimate (M-step) obtained by E-step, re-estimate and update the parameters, and then start the next round of E-step calculation, repeat this EM process until the parameters converge [16].

III. Feature Recognition Method of Residential User Data based on Semi-supervised Learning

A. Overall Methodology Framework

The process of identifying the features of residential user data based on semi-supervised learning algorithm can be divided into three parts, and the overall method framework is shown in Fig. 1.

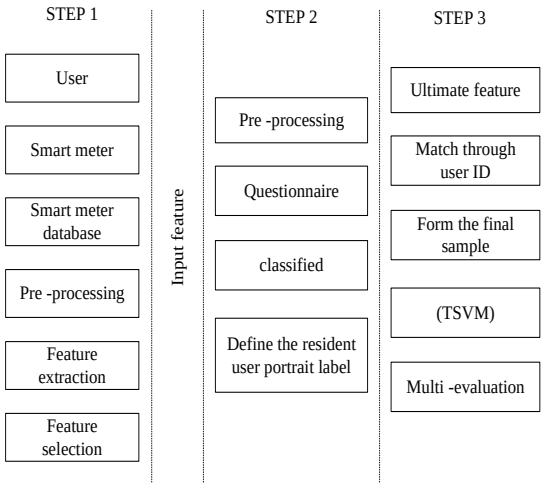


Fig. 1. Overall Method Framework of Resident user Data Feature Recognition Model based on Semi-supervised Learning.

1) First, remove the abnormal values and missing values of the smart meter data, then extract 54 time domain features and 24 frequency domain features. Normalize the features, and finally use the feature selection algorithm to screen the extracted load features, and select the input features of the subsequent identification model according to the importance ranking [17].

2) Sort out the questionnaire data related to user data characteristics, mainly analyze the user data characteristics that have a greater impact on power consumption, and classify and define the classification indicators for each type of user data characteristics according to the answer results of the questionnaire. The classification indicators can be used as real user indicators for the calculation of model performance indicators [18].

3) Match the finally selected smart meter data characteristics with the resident user data characteristic indexes through the user ID to form a final sample set. The semi-supervised learning method is used to train a small number of samples with indicators and a large number of samples without indicators to obtain the user data feature recognition model based on semi-supervised learning. Finally, the performance evaluation index is used to verify the effectiveness of the model [19].

B. Feature Selection

In this paper, three methods are used to select features respectively to facilitate the subsequent verification of the impact of different feature selection methods on semi-supervised learning to identify the features of residential user data.

1) Filtration: For the 78 data features extracted from the CER data set, the filtering method first uses the variance discrimination method, sets the variance threshold, and eliminates the features whose variance is less than the threshold (that is, there is no discrimination) [20]. Then the Pearson correlation coefficient method is used to calculate the Pearson correlation coefficient of the remaining features and the real user data feature identification classification index, and the R most important (i.e., most relevant) features are selected according to the coefficient value ranking [21].

2) Packaging method: For the packaging method, the Recursive Feature Elimination (RFE) method based on Logistic Regression (LR) algorithm is used in this section to select the features extracted from the CER data set. Firstly, a weight value is assigned to each original feature in the initial training. Secondly, the LR model is used to predict the classification index. Then the predicted classification index is compared with the real index to calculate the recognition error. The weight of each feature is updated according to the error, and the feature with the smallest absolute value of the weight is proposed in each round [22]. Repeat this step until the required number of features is reached. Finally, these features with larger weights are used as the input of the subsequent classification model. A schematic diagram of feature selection based on the LR-RFE algorithm is shown in Fig. 2.

3) Embedding method: The Random forest measures its ability to identify user data features by calculating the PI value of each feature. When calculating the importance of the extracted feature N, a decision tree i is created. The OOBErrori is first calculated. Then, the values of the out-of-bag data feature N are randomly rearranged, and the rest of the features remain unchanged to form a new out-of-bag data set OOBi. According to the new OOB, the OOBErrori is recalculated. The PI value of the feature N in the ith tree can be obtained by subtracting the results of the two calculations, and the calculation formula is shown in (3).

$$PI_i(N) = OBError_i^{'} - OOBError_i \tag{3}$$

The calculation process is repeated for each tree of the random forest, and the final PI value of the feature N can be obtained by summing and averaging the PI values of the feature N of each tree, as shown in (4).

$$PI(N) = \frac{1}{c}\sum_{i=1}^{c} PI_i(N) \tag{4}$$

C = ntree represents the number of decision trees used in the random forest. If the importance of a feature ranks high, it means that its value has discrimination between different samples. After the eigenvalues are randomly reordered on the out-of-bag data set, their discrimination for different user samples is reduced, thus improving the OOBErrori. Therefore, the higher the PI value, the higher the importance of the feature. The process of using the random forest algorithm to rank the importance of features is shown in Fig. 3.

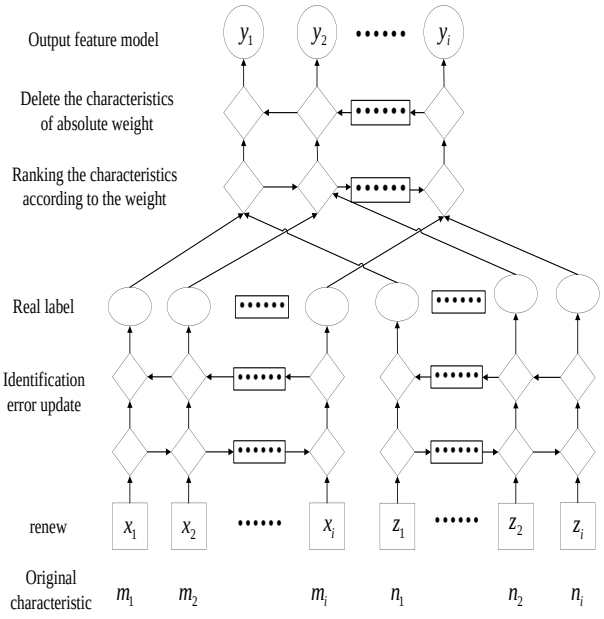


Fig. 2. Schematic Diagram of Feature Selection based on LR-RFE Algorithm.

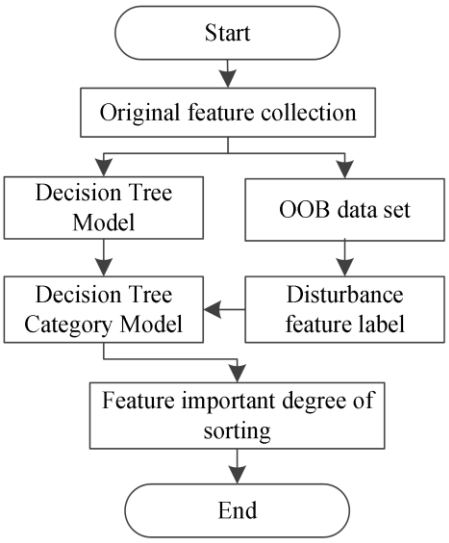


Fig. 3. Ranking Process of Feature Importance Calculated by Random Forest Algorithm.

C. Performance Evaluation Index

In this paper, the TSVM semi-supervised learning method is used to identify the characteristics of residential user data. Three performance evaluation indicators, namely ACC, F1-Score and AUC, are used to evaluate and analyze the recognition model.

When training the SVM classifier, five-fold cross validation is used to train the samples to verify the recognition performance of the data features of residential users more reliably. To evaluate the classification performance of the classifier from multiple perspectives, several evaluation indexes are given here.

- Accuracy

For user data features with M category indexes, the confusion matrix C of $M \times M$ can be calculated. J represents the number of features with the category index of m that are misclassified into the category index of n. If m = n, then Cm, n represent the number of correct classifications, and vice versa. Accuracy (ACC) can be expressed by formula (5).

$$Accuracy = \frac{\sum_{m=1}^{M} C_{m,m}}{\sum_{m=1}^{M}\sum_{n=1}^{M} C_{m,m}} \tag{5}$$

For the binary classification problem, the confusion matrix shown in Table I can be obtained by comparing the sample indicators identified by the classification model with the real sample indicators.

TABLE I. BINARY CLASSIFICATION CONFUSION MATRIX

	Positive	Negative
The prediction is positive	TP	FP
The prediction is negative	FN	TN

True Positives (TP): the number of samples that are actually positive and predicted to be positive; False Positives (FP): the number of samples that are actually negative and predicted to be positive.

False Negatives (FN): the number of samples that are actually negative and predicted to be positive.

True Negatives (TN): the number of samples that are actually negative and predicted to be negative. Therefore, ACC can also be expressed by the following formula (6):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

- F1-Score

According to the confusion matrix, several performance evaluation indexes can be obtained.

Precision: The proportion of samples that are actually positive (positive) among the samples predicted to be positive (positive), as shown in formula (7).

$$Presicion = \frac{TP}{TP + FP} \tag{7}$$

Recall: The proportion of samples that are correctly classified as positive among all samples that are really positive (positive examples), as shown in formula (8).

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{8}$$

F1-Score is a comprehensive index reflecting precision and recall, and its value range is 0 to 1. The closer the value of F1-Score is to 1, the better the recognition performance of the model is, and its calculation formula is shown in (9).

$$F1 - Score = \frac{2 \times Presicion \times Recall}{Presicion + Recall} \tag{9}$$

IV. EXPERIMENTAL ANALYSIS

A. Experimental Environment Setting

This paper also uses the open CER data set to analyze the data characteristics of six residential users. See Table II for details.

To verify the effectiveness of the proposed method, the recognition performance of TSVM semi-supervised learning algorithm is compared with four classical KNN, RF, SVM and MLP supervised learning algorithms, and two examples are set to verify the results.

To verify whether the recognition performance of the resident user portraits can be improved by adding frequency domain features on the basis of time domain features, SVM is used to train the features of smart meters, and the grid search method is used to optimize the SVM parameters. 80% of the user sample is the training set and 20% is the test set. The selection of the main hyper-parameters of SVM in the three cases of only time domain feature (Model 1), only frequency domain feature (Model 2), and combination of time domain and frequency domain (Proposed Model) is shown in Table III.

TABLE II. USER DATA CHARACTERISTIC DESCRIPTION AND INDEX DEFINITION TABLE BASED ON CER DATA SET

Number	User data Characteristics	User data characterization	Category	Indicators	Sample size
1	Employment	Employment status of the family's main earner	Hire	1	1423
			Not hired	2	1026
2	Population	Number of family members	Little (no<2)	1	1321
			Many (no.N3)	2	1128
3	Housing types	Housing types	Freestyle	1	1299
			Connection type	2	1104
4	Occupancy rate	Is the house unused for more than 6 hours per day?	Yes	1	1619
		Is the house unused for more than 6 hours per day?	No	2	345
5	Cooking type	Type of cooking facility	Electricity	1	1712
			Non-electricity	2	737
6	With or without children at home	With or without children at home	Yes	1	1964
			No	2	485

TABLE III. MAIN HYPER-PARAMETER SETTINGS OF SVM CLASSIFIERS

SVM	Model 1				Model 2			Proposed model	
serial number	Kernel	C	gamma	Kernel	C	gamma	Kernel	C	gamma
1	RBF	99	0.02	RBF	23	0.2	RBF	5	0.02
2	RBF	97	0.02	RBF	85	0.2	RBF	59	0.002
3	RBF	33	0.02	RBF	3	0.2	RBF	67	0.02
4	RBF	2	20	RBF	2	200	RBF	2	20
5	RBF	22	0.02	RBF	43	20	RBF	32	0.02
6	RBF	2	0.2	RBF	77	0.2	RBF	2	0.2
7	RBF	2	20	RBF	2	200	RBF	2	20

First of all, for the basic results of TSVM, the accuracy values of population, housing occupancy, cooking type and whether there are children in the house are higher than 75%, and the accuracy values of employment and housing type are between 60% and 70%. The recognition accuracy of all user data features is higher than 60%, which shows the basic effectiveness of the semi-supervised learning method proposed in this paper.

Comparing the TSVM semi-supervised learning algorithm with other supervised learning algorithms, when the proportion of index samples is set to 5%, for any user data feature, the recognition accuracy, F1-Score and AUC of TSVM are higher than those of other supervised learning algorithms.

B. Example 2 Experimental Setup and Result Analysis

To further verify the performance of the proposed method, the proportion of samples with index for the semi-supervised learning algorithm is set to 5%, and the proportion of samples with index for the supervised learning algorithm is set to 10 times that of the semi-supervised learning model, that is, 50%. Here, the LR-RFE-based feature selection method is still used to select the top 20 features that are correlated with the classification target value. In this case, the ACC, F1-Score and AUC values of the TSVM semi-supervised learning algorithm and the four supervised learning algorithms of KNN, RF, SVM and MLP to identify the characteristic indicators of residential user data are shown in Fig. 4.

In Fig. 4, for the TSVM semi-supervised algorithm, except for the user-data feature of population number, the recognition ACC values of other user data features are higher than those of other supervised learning methods. For the population, the ACC and AUC values identified by TSVM with only 5% of the matching data are slightly lower than those identified by the supervised learning method with 50% of the matching data, but the difference is not significant.

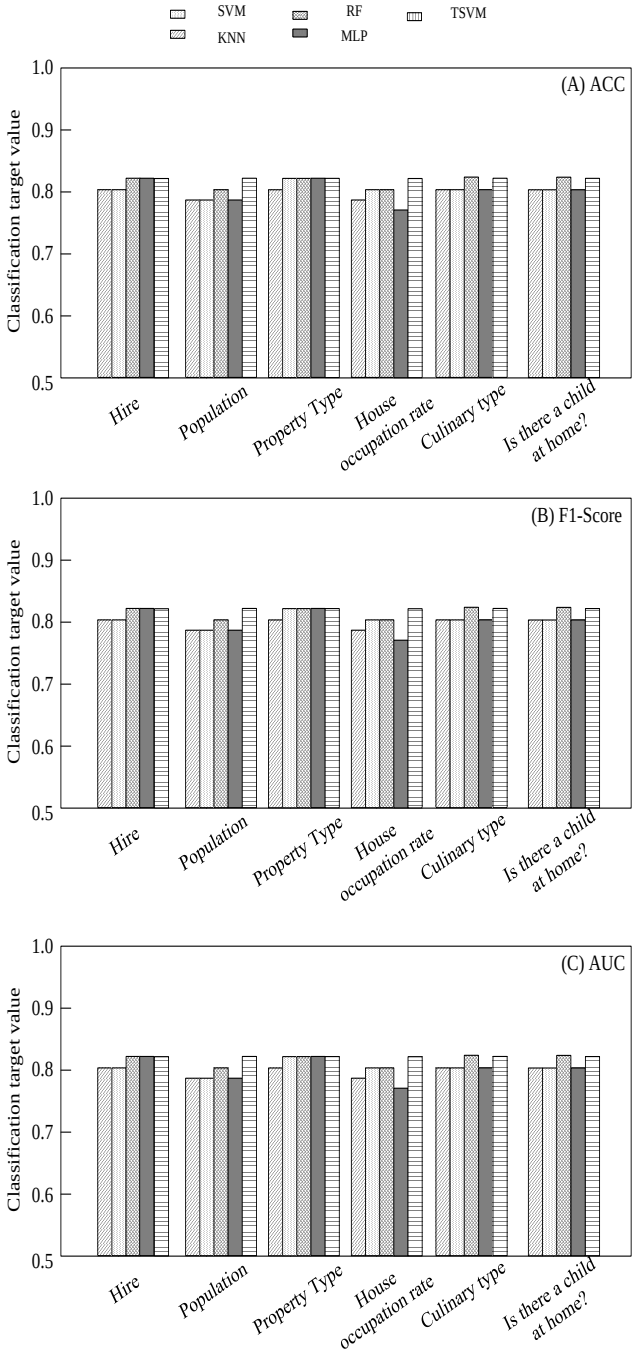


Fig. 4. Comparison of Semi-supervised and Supervised Recognition Algorithms.

V. CONCLUSION

In this paper, a feature recognition method for residential user data based on semi-supervised learning in the case of limited index samples is proposed. The main work includes:

1) Based on the extracted time domain and frequency domain features of smart meters, a feature recognition method for residential user data based on semi-supervised learning is proposed.

2) Then, the effectiveness of semi-supervised learning is verified by the real CER data set.

3) Two main factors affecting the performance of semi-supervised learning recognition algorithms are analyzed.

This method can make full use of the potential rules contained in a small number of matching data and a large number of non-index data to explore the relationship between smart meter data and residential user data characteristics and reduce the cost of index marking.

In future work, it is necessary to further explore the relationship between smart meter data and residential electricity consumption behavior habits, update the identification scenario, and integrate the user data characteristic information into residential load pattern forecasting, baseline load estimation or high demand response potential user identification from the user demand side.

REFERENCES

[1] Meher S. Semi-supervised self-learning granular neural networks for remote sensing image classification. Applied Soft Computing, 2019, 83:105655.

[2] Muhammad F, Alberto S. Analyzing load profiles of energy consumption to infer household characteristics using smart meters. Energies, 2019, 12(5):773.

[3] Viegas J L, Vieira S M, Melício R, et al. Classification of new electricity customers based on surveys and smart metering data. Energy, 2016, 107:804-817.

[4] Zhong S, Tam K S. Hierarchical classification of load profiles based on their characteristic attributes in frequency domain. IEEE Transactions on Power Systems, 2015, 30(5):2434–2441.

[5] Gajowniczek K, Ząbkowski T, Sodenkamp M. Revealing household characteristics from electricity meter data with grade analysis and machine learning algorithms. Applied sciences, 2018, 8(9):1654.

[6] Hopf K, Sodenkamp M, Kozlovkiy I, et al. Feature extraction and filtering for household classification based on smart electricity meter data. Computer Science Research and Development, 2016, 31(3):141–148.

[7] Sun G, Cong Y, Hou D, et al. Joint household characteristic prediction via smart meter data. IEEE Transactions on Smart Grid, 2017, 10(2):1834-1844.

[8] Wang Y, Chen Q, Gan D, et al. Deep learning-based socio-demographic information identification from smart meter data. IEEE Transactions on Smart Grid, 2019, 10(3):2593-2602.

[9] Wang Y, Bennani I, Liu X, et al. Electricity consumer characteristics identification: a federated learning approach. IEEE Transactions on Smart Grid, 2021, Early Access.

[10] Wang F, Li K, Duić N, et al. Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. Energy Conversion and Management, 2018, 171:839-854.

[11] Haben S, Singleton C, Grindrod P. Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data. IEEE Transactions on Smart Grid, 2017, 7(1):136-144.

[12] Kumar P, Banerjee R, Mishra T. A framework for analyzing trade-offs in cost and emissions in power sector. Energy, 2020, 195:116949.

[13] Li K, Cao X, Ge X, et al. Meta-heuristic optimization based two-stage residential load pattern clustering approach considering intra-cluster compactness and inter-cluster separation. IEEE Transactions on Industry Applications, 2020, 56(4):3375-3384.

[14] Wang F, Li K, Liu C, et al. Synchronous Pattern Matching Principle-Based Residential Demand Response Baseline Estimation: Mechanism Analysis and Approach Description. IEEE Transactions on Smart Grid, 2020, 9(6):6972-6985.

[15] Li K, Wang F, Mi Z, et al. Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation. Applied Energy, 2019, 253:113595.

[16] Wang F, Xiang B, Li K, et al. Smart households' aggregated capacity forecasting for load aggregators under incentive-based demand response programs. IEEE Transactions on Industry Applications, 2020, 56(2):1086-1097.

[17] Wang F, Ge X, Yang P, et al. Day-ahead optimal bidding and scheduling strategies for DER aggregator considering responsive uncertainty under real-time pricing. Energy, 2020, 213:118765.

[18] Lu Q, Lü S, Leng Y, et al. Optimal household energy management based on smart residential energy hub considering uncertain behaviors. Energy, 2020, 195:117052.

[19] Finck C, Li R, Zeiler W. Economic model predictive control for demand flexibility of a residential building. Energy, 2019, 176:365-379.

[20] Li K, Mu Q, Wang F, et al. A business model incorporating harmonic control as a value-added service for utility-owned electricity retailers. IEEE Transactions on Industry Applications, 2019, 55(5):4441-4450.

[21] Li K, Liu L, Wang F, et al. Impact factors analysis on the probability characterized effects of time of use demand response tariffs using association rule mining method. Energy Conversion and Management, 2019, 197:111891.

[22] Lin J, Marshall K R, Kabaca S, et al. Energy affordability in practice: Oracle Utilities Opower's business Intelligence to meet low and moderate income need at Eversource. The Electricity Journal, 2020, 33(2):106687.

Light Gradient Boosting with Hyper Parameter Tuning Optimization for COVID-19 Prediction

Ferda Ernawan[1]
Faculty of Computing
Universiti Malaysia Pahang
Pekan, Malaysia

Kartika Handayani[2]
Faculty of Engineering and Informatics
Universitas Bina Sarana Informatika
Jakarta, Indonesia

Mohammad Fakhreldin[3]
Faculty of Computer Science and Information Technology
Jazan University
Jazan, Saudi Arabia

Yagoub Abbker[4]
Faculty of Computer Science & Information Technology
Jazan University
Jazan, Saudi Arabia

***Abstract*—The 2019 coronavirus disease (COVID-19) caused pandemic and a huge number of deaths in the world. COVID-19 screening is needed to identify suspected positive COVID-19 or not and it can reduce the spread of COVID-19. The polymerase chain reaction (PCR) test for COVID-19 is a test that analyzes the respiratory specimen. The blood test also can be used to show people who have been infected with SARS-CoV-2. In addition, age parameters also contribute to the susceptibility of COVID-19 transmission. This paper presents the extra trees classification with random over-sampling by considering blood and age parameters for COVID-19 screening. This research proposes enhanced preprocessing data by using KNN Imputer to handle large missing values. The experiments evaluated the existing classification methods such as Random Forest, Extra Trees, Ada Boost, Gradient Boosting, and the proposed Light Gradient Boosting with hyperparameter tuning to measure the predictions of patients infected with SARS-CoV-2. The experiments used Albert Einstein Hospital test data in Brazil that consisted of 5,644 sample data from 559 patients with infected SARS-CoV-2. The experimental results show that the proposed scheme achieves an accuracy of about 98,58%, recall of 98,58%, the precision of 98,61%, F1-Score of 98,61%, and AUC of 0,9682.**

***Keywords*—ROS; light gradient boosting; hyper parameter tuning; COVID-19 screening; blood and age based**

I. Introduction

Coronavirus 19 (COVID-19) is a highly contagious viral infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. SARS-CoV-2 can cause tissue damage and cause acute respiratory distress syndrome. It is rapidly increasing transmission rate which demands an early response to diagnose and prevent the rapid spread of this disease [2]. Currently, COVID-19 is being transmitted by human-to-human through air transmission that cause a wide spread of the disease [3]. One way to detect COVID-19 is through the Reverse-Transcriptase Polymerase Chain Reaction, also known as RT-PCR [4]. RT-PCR has limited resources, it has high specificity and high sensitivity [5]. However, according to the study of validation of the SARS-CoV-2 RT-PCR test [6], blood or hematological parameters showed high sensitivity and specificity as well as intra and inter-test precision and efficiency.

Machine learning can become an alternative for diagnosing and analyzing COVID-19 infection [7]. Machine Learning has been widely used to investigate and help in screening with suspected COVID-19 infection [8]. The implementation of machine learning in RT-PCR with blood assessments has a critical function for diagnosing COVID-19 and different respiration diseases. The parameters are involved white blood cells, C-reactive protein, neutrophils, lymphocytes, monocytes, eosinophils, basophils, aspartate and alanine, lactate dehydrogenase, and others. Those parameters have proven an excessive correlation in sufferers identified with COVID [9]. In addition, age parameters [10] also affect the susceptibility of COVID-19 transmission. Therefore, it motivates researchers to investigate parameters that significantly effect for covid-19 prediction.

This research presents a predictive model for diagnosing COVID-19 by considering C-reactive protein, neutrophils, lymphocytes, monocytes, eosinophils, basophils, aspartate and alanine, lactate dehydrogenase, including blood and age parameters. This research proposes a predictive model by using ensemble learning which involved Random Forest, Extra Trees, AdaBoost, Gradient Boosting and Light Gradient Boosting, then optimizes the best model with hyperparameter tuning. The experiments also investigate the best solution for imbalance data by implementing the existing sampling methods such as Random Under Sampling (RUS), Random Over Sampling (ROS) and Synthetic Minority Over Sampling TEchnique (SMOTE). The sampling class imbalance approaches is used to overcome imbalance data that has been carried out in the research related to Covid-19 [11]. This research is expected to obtain the best predictive model that can achieve high accuracy, recall, precision, f-score and AUC compared to the existing schemes.

II. Related Work

Several researches have proven the significant of blood exams for the diagnosis of Covid-19 [12] analyzing the blood

index of 69 COVID-19 sufferers. All have been dealt with on the National Center for Infectious Diseases (NCID) placed in Singapore. Among those sufferers, sixty-five underwent whole blood assume the day of admission. In addition, demographic facts inclusive of age, gender, ethnicity, and region have been furnished for this study. Around 13,4% of sufferers require in-depth care unit (ICU) care, specifically the elderly. During the primary examination, 19 sufferers had leukopenia (low white blood cells) and 24 had lymphopenia (low lymphocyte stage with inside the blood), with five instances categorized as severe (Absolute lymphocyte count (ALC).

The application of a Covid-19 diagnosis based on blood tests has previously been carried out to provide comprehensible answers primarily based totally on device studying techniques using public data from the Albert Einstein Hospital. Previously, data preprocessing was carried out for selection of blood features. Then normalization of features with z-score and use of iterative imputer method to fill in missing values is done. The remaining 608 patients, 84 of whom have been high-quality for COVID-19 showed with the aid of using RT-PCR [13]. In order to apprehend the decisions, a neighborhood Decision Tree Explainer (DTX) approach is performed to obtain the results.

Data from the Israel Albert Einstein Hospital located in São Paulo, Brazil are also used in the application of machine learning in the diagnosis of COVID-19 with hematological parameters. Pre-processing is done by selecting features using particle swarm optimization (PSO) and evolutionary search (ES). Furthermore, experiments were carried out with different machine learning techniques. The experimental results show that Bayesian networks [7] have superior performance compared to other techniques with an overall accuracy of 95,159%, kappa index 0,903, sensitivity 0,968, precision 0.938, and specificity 0,936.

A study was also conducted to identify SARS-CoV-2 positive patients from a total of 598 complete data and 5046 were not used because they were incomplete. A machine learning model, ANN was carried out to test based on the dataset obtained from the Israelta Albert Einstein Hospital, in São Paulo, Brazil by testing various hematological parameters. As a result, the flexible ANN model [14] predicts COVID-19 patients with high accuracy between the population in the regular ward AUC 94-95% and those not hospitalized or in the community AUC 80-86%.

Other research was conducted by building a two-stage test; in level one, no preprocessing technique is carried out even as in level preprocessing is emphasized to attain higher predictive effects. Blood samples from sufferers from Einstein Hospital in Brazil were amassed and used for prediction of the severity of COVID-19 with studying algorithms. The Tuned Random Forest algorithm [15] produced an accuracy of 0,98 with numerous preprocessing methods.

Based on the description of the related research above, the existing considers few parameters to diagnose COVID-19. There are a quite few research studies on blood exams for the diagnosis of COVID-19. However, studies on eosinophils, age and blood parameters are rare to find in literature. This study proposes a pre-processing KNN imputer data to overcome the large missing values. Then various data sampling class with imbalance approaches methods is used to find out the best sampling class for imbalance datasets. Whereas the prediction model generated from the data classification process using an ensemble, namely Extra Trees, Bagging Decision Tree, Random Forest, Ada Boost, Gradient Boosting and Light Gradient Boosting.

III. Preliminaries

A. Ensemble Learning Classification Model

Ensembles learning classification model can increase the computational costs [16], as it is necessary to train several individual classifiers, and their computational requirements can grow exponentially when dealing with large scales.

B. Extra Trees

The extra tree classifier creates a gaggle of unpruned decision trees in step with the standard top-down method. The predictions of all trees were combined to determine the ultimate prediction, through the majority alternative [17]. The extra tree classifier generates a random multiple of the choice tree with completely different sub-samples while not bootstrapping. The extra trees can avoid over-fitting issues and improves accuracy [18]. Efficiency is also the main strength of this study.

C. AdaBoost

AdaBoost is an iterative algorithm, in each iteration, instances that were wrongly classified in the previous iteration are given more weight. Sequentially apply the learning algorithm to reweighted the sample from the original training data. Initially, each instance is ssigned the same weight and iteration as the iteration, the weight of all misclassified instances is increased and the correctly classified instances are reduced [17]. The AdaBost algorithms [19] are defined by:

1) Minimize the error function with the formula

$$\mathrm{w_e} = \sum_{yi \neq km(xi)} w_i^{(m)} \exp(a_m) \tag{1}$$

2) Set the value a with the formula

$$a_m = \frac{1}{2} n\left(\frac{1-e_m}{e_m}\right), e_m = \frac{w_e}{w} \tag{2}$$

3) Update values if observing misclassification by formula

$$w_i^{(m+1)} = w_i^{(m)} \exp(a_m) = w_i^{(m)} \sqrt{\frac{1-e_m}{e_m}} \tag{3}$$

4) For other values using the formula

$$w_i^{(m+1)} = w_i^{(m)} \exp(a_m) = w_i^{(m)} \sqrt{\frac{e_m}{1-e_m}} \tag{4}$$

D. Gradient Boosting

Gradient Boosting is a machine learning algorithm that can solve regression and classification problems. Gradient Boosting generates a prediction model consisting of an ensemble of weak prediction models in the decision tree [20]. The construct of a gradient boosting call tree is to mix a series of weak base classifiers into one sturdy one. a conventional boosting methodology that weighs positive and negative samples, GBDT builds a world convergence rule by following the direction of the negative gradient [21]. The GBDT measures GBDT [21] are presented as follows.

1) Step 1: The values for the initial constants of the model β are given:

$$f0(x) = \arg\min \beta \sum_{i=1}^{n} L(yi, \beta) \tag{5}$$

2) Step 2: For the number of iterations m = 1: M (M is the iteration time), the residual gradient direction is calculated

$$y_i - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right] f(x) - f_{m-1(x)} \tag{6}$$

3) Step 3: Base classifiers are used to adjust the sample data and obtain the initial model. According to the least squares approach, the parameters of the model are obtained and the model *h (xi; am*) is installed

$$a_m = \arg\min_{a\beta} \sum_{i=1}^{N} \left[y_i^* - \beta h(x_i; a) \right]^2 \tag{7}$$

4) Step 4: Function loss is minimized. According to Eq. (4), the new step size of the model, i.e. the weight of the current model, is calculated.

$$\beta_m = \arg\min_{a\beta} \sum_{i=1}^{N} L\left[y_i^*, f_{m-}(x) = +\beta h(x_i; a) \right] \tag{8}$$

5) Step 5: the model is updated as follows

$$F_m(x) = F_{m-1}(x) + \beta_m h(x_i; a) \tag{9}$$

E. Light Gradient Boosting

Light Gradient Boosting Machine or LightGBM uses gradient enhancement in its construction, but light GBM does not divide the eigenvalues one by one, so it is necessary to calculate the splitting benefit of each eigenvalue. LightGBM algorithm on the model to improve forecasting accuracy and robustness [22]. It can indeed find the optimal split value, but it costs a lot, and may not be good for generalizing information when the amount of data is large [23]. Remembering the supervised training $setX = \{(xi, yi)\}_i^n$ LightGBM's target is to find approximation for a particular function $f(x)$ to a certain function $\hat{f}(x)$ which reduces the expected loss function value , $(y, f(x))$ as follows [24] :

$$\hat{f} = \arg\min Ey, xL(y, f(x)) \tag{10}$$

LightGBM integrates a number of T regression trees to approach the final model, which is.

$$f_T(x) \sum_{t=2}^{T} f_t(x) \tag{11}$$

$q(x), q \in \{1,2...., J\}$, where J denotes the number of leaves, represents the guideline of thumb of the choice tree and is the leaf node weight vector. Therefore, LightGBM could be educated additively inside the following steps:

$$T_t = \sum_{i=1}^{n} L(yi, f_{t-1}(xi) + f_t(xi)) \tag{12}$$

In LightGBM, Newton' technique simply approximates the target function. Where *gi* and *hi* indicate the first- and second-order gradient statistics of the loss function, let *Ij* show the instance set of leaf*j*.

$$T_t = \sum_{i=1}^{n} \left(\left(\sum_{i \in Ij} g_i \right) + \frac{1}{2} \left(\sum_{i \in Ij} h_i + \lambda \right) w_j^2 \right) \tag{13}$$

For the tree structure$q(x)$, the optimum leaf weight score of every leaf node $w*$ and therefore the extreme worth of T_t may be solved as follows:

$$w_j^* = -\frac{\sum_{i \in Ijgi}}{\sum_{i \in Ij} (h_i + \lambda) w_j^2} \tag{14}$$

F. Random Forest

Random Forest is an integrated learning method based on bagging. The essence is to apply the bootstrap method to the CART algorithm. Random Forest samples were taken using the bootstrap method, and then an independent decision tree model was built using the CART algorithm [25]. Random forest algorithm (for each type and regression) [26] are discussed as follows:

1) From Training n samples draw n_{tree} bootstrap samples.

2) For every of the bootstrap samples, develop classification or regression tree with the subsequent modification: at every node, in place of selecting the excellent break up amongst all predictors, randomly pattern m_{trv} of the predictors and select the excellent break up amongst the ones variables. The tree is grown to the most length and not pruned back. Bagging may be concept of because the unique case of random forests received while $m_{try} = p$, the wide variety of predictors.

3) Predict new facts by combining n_{tree} tree predictions (i.e., majority vote for type, common for regression.

G. Random Over Sampling (ROS)

ROS algorithm randomly replicates samples from the minority classes [27]. Oversampling [28] can be done by

increasing number of instances or minority class samples by production new instance or repeated multiple instances.

H. Random Under Sampling (RUS)

RUS technique at random eliminates samples from the bulk categories, till achieving a relative categories balance [27]. For the under-sampling approach, most of the category instances are discarded till additional a balanced distribution of information is achieved. This data merchandising method is completed every which way. Considering an information set with a hundred minority class instances and 2,000 magnitude class instances, a complete of 1800 categories that majority are going to be deleted randomly within the RUS technique. The dataset will be balanced with two hundred instances, it will be delineating with 200 instances, whereas minority also have 200.

I. Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE produces artificial samples from the minority class by interpolating existing instances that are terribly near to every other [27]. For the minority category within the information set, SMOTE initial selects the minority class data instance randomly. The distance from the sample set to several classes is calculated by the Euclidean distance *D*, and K-nearest neighbors are obtained. The Euclidean distance *D* is defined by:

$$D = \sqrt{\sum_{k-1}^{n} (X_k - Y_k)^2} \tag{15}$$

According to the proportion of the unbalanced data set, the sampling rate N is set. The six samples closest to D were selected as one group. Each sample group is connected to each other to generate several new samples at random, which are added to the data set and recycled [29]. This results in a new formula:

$$X_{new} = x_i + rand(0,1) * |x_i - x_j| \tag{16}$$

IV. Experimental Setup

Images are divided by 70% for training, 20% for validation, and 10% for testing. Then the YOLO architectural model is used from training and validation and then a data test is carried out with data testing and detecting disease. After that, a performance evaluation's carried out for the architectural model used. The block diagram of the proposed covid-19 classification is depicted in Fig. 1.

This study uses machine learning techniques to predict negative and positive cases using RT-PCR data with blood parameters. Before applying the machine learning classification method, data preparation was carried out by using several methods, namely, Remove non-blood parameter, Imputation Missing Values, Label Encoding Class and Normalization with Z-Score. The processed data was tested using several machine learning classification methods using an ensemble, namely Extra Trees, Bagging Decision Tree, Random Forest, Ada Boost, Gradient Boosting and Light Gradient Boosting. In testing the machine learning classification method, the best method was chosen based on the evaluation of the results in terms of accuracy, precision, recall, F-1score and AUC. The best method is optimized by searching for the best parameters by using hyper parameter tuning. Then, the results were compared before using hyper parameter tuning and after using hyper parameter tuning. The results of the best methods can be used for prediction of COVID-19.

A. Data Collection

The dataset is collected from the existing benchmark [30]. The dataset consists of 5644 patients treated at the Albert Einstein Israelta Hospital located in Saulo Paulo, Brazil. Kaggle makes data sets available for public access. Data was collected from 28 March 2020 to 3 April 2020, with more than 100 laboratory tests including blood test, urine test, SARS-CoV-2 test, RT-PCR test, presence of influenza virus [30]. The dataset consists of 89% missing values, so the missing value is handled by filling in the missing value using the KNN Imputer method using K = 5 [31]. Label encoding is done which aims to perform coding on the class label. Label Encoding serves to change the data format of numbers 0 to n_classes-1, this is intended to make data training easier. Normalization of the data was performed using Z-Score [32]. Then the best method is to optimize hyper parameter tuning using GridSearchCV. GridSearchCV taken from Scikit learn [33]. This study considers several features for classification as shown in Table I.

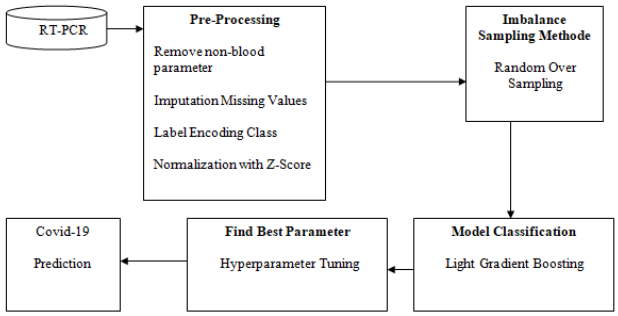


Fig. 1. Block Diagram of Covid-19 Prediction.

TABLE I. Selection of Features

No.	Features	No.	Features
1	Hematocrit	13	Red blood cell distribution width
2	Hemoglobin	14	Monocytes
3	Platelets	15	Mean platelet volume
4	Red blood Cells	16	Neutrophils
5	Lymphocytes	17	C-reactive protein
6	Mean corpuscular hemoglobin	18	Creatinine
7	MCH concentration	19	Urea
8	Leukocytes	20	Potassium
9	Basophils	21	Sodium
10	Eosinophils	22	Aspartate transaminase
11	Lactate dehydrogenase	23	Alanine transaminase
12	Mean corpuscular volume	24	Age

B. Split Validation

In this study, the experiments divide the data based on the ratio entered, for example the percentage of 80:20 [34]. There are 80% of the total amount for training set and 20% for test set.

C. Evaluation

To compare the overall performance of the proposed scheme, we decided on five metrics: accuracy, recall, precision, F1-Score and receiver running characteristic (ROC) curves, and the cost of the vicinity below the ROC curve (ROC AUC). Accuracy is the maximum generally used assessment metric for type. However, for imbalance facts type problems, accuracy won't be a great preference due to the fact accuracy regularly has a bias closer to the bulk class [35][36]. The accuracy can be defined by:

$$Accuracy = \frac{TP + TN}{TotalSample} \tag{17}$$

Recall is the collection of data that has been successfully taken from the part of the data relevant to the query [37]. The Recall is defined by:

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{18}$$

Precision is part of the data taken in accordance with the required information [38]-[40]. The precision is defined by:

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \tag{19}$$

The F1 score is the Harmonic Mean between precision and Recall [41]. The F-Score indicates how precise the classifier is (how many instances are correctly classified), as well as how strong it is (it doesn't miss a large number of instances). The F1-Score formula is defined by:

$$F1 - Score = 2\frac{\mathrm{Re}call * \mathrm{Pr}ecision}{\mathrm{Re}call + \mathrm{Pr}ecision} \tag{20}$$

The ROC curve represents the genuine advantageous rate (TPR) and fake advantageous rate (FPR). TPR represents the ratio of advantageous samples that have been successfully detected through the algorithm, and FPR represents the ratio of terrible samples that have been incorrectly labeled as advantageous. The expressions for TPR and FPR are as follows:

$$TPR = \frac{TP}{TP + FN} \tag{21}$$

$$FPR = \frac{FP}{TN + FP} \tag{22}$$

where TP is the number of true positives, TN is the number of true negatives, FN is the number of false negatives, FP is number of false positives.

V. Experimental Results

After pre-processing the data to overcome the missing value, performing a Z-Score then encoding the dataset class, testing the specified model without using the sampling class imbalance approaches method. Testing the model without sampling class imbalance approaches method is carried out first for further comparison with various sampling class imbalance approaches methods to be tested. The test results are listed in Table II.

The best accuracy was obtained by using extra trees method with an average accuracy of 98.40% for imbalance sampling method. While, the light gradient boosting achieved high accuracy with random under sampling than extra trees, AdaBoost, Gradient Boosting, and Random Forest methods. Overall, the extra trees method performs better than other method for different types of sampling method except random under sampling. The experimental results in terms of recall, precision, F1-Score, and AUC are listed in Tables III, IV, V and VI.

The classification of light gradient boosting method achieved recall value of 91.96%. The best recall result was obtained from sampling technique of without imbalance sampling method, random under sampling, SMOTE and SMOTE-Tomek. The experiments also evaluate the precision of the classification method; classification by using extra tree produced a high precision result except sampling technique of random under sampling. The classification of light gradient boosting method can achieve a good F1-Score and AUC score under various sampling techniques. The visual comparison of the accuracy, recall, precision, F1-score and AUC is shown in Fig. 2, 3, 4, 5 and 6.

TABLE II. Selection of Accuracy Result for 5644 RT-PCR Data

Sampling Technique	Extra Trees	Light Gradient Boosting	AdaBoost	Gradient Boosting	Random Forest
Without imbalance sampling method	**98,4**	98,22	96,89	97,69	98,22
Random Under Sampling	96,27	**96,63**	95,3	96,27	96,63
Random Over Sampling	**98,4**	98,22	96,89	97,69	98,22
SMOTE	**98,4**	98,22	97,34	97,6	97,96
SMOTE- Tomek	**98,49**	98,22	97,34	97,69	98,05

TABLE III. SELECTION OF RECALL RESULT FOR 5644 RT-PCR DATA

Sampling Technique	Extra Trees	Light Gradient Boosting	AdaBoost	Gradient Boosting	Random Forest
Without imbalance sampling method	88,39	**91,96**	86,60	90,17	89,28
Random Under Sampling	96,42	**97,32**	94,64	93,75	96,42
Random Over Sampling	88,39	91,96	**94,64**	93,75	90,17
SMOTE	90,17	**91,96**	90,17	89,28	88,39
SMOTE- Tomek	90,17	**91,96**	90,17	89,28	88,39

TABLE IV. SELECTION OF PRECISION RESULT FOR 5644 RT-PCR DATA

Sampling Technique	Extra Trees	Light Gradient Boosting	AdaBoost	Gradient Boosting	Random Forest
Without imbalance sampling method	**95,19**	94,49	90,65	92,66	93,45
Random Under Sampling	72,97	75,69	69,28	75	**76,05**
Random Over Sampling	**95,19**	90,35	78,51	84,67	91,81
SMOTE	**91,81**	91,15	87,06	88,49	89,59
SMOTE- Tomek	**94,39**	90,35	84,16	87,71	91,66

TABLE V. SELECTION OF F1-SCORE RESULT FOR 5644 RT-PCR DATA

Sampling Technique	Extra Trees	Light Gradient Boosting	AdaBoost	Gradient Boosting	Random Forest
Without imbalance sampling method	91,66	**93,21**	88,58	91,4	91,32
Random Under Sampling	83,07	**85,15**	80	83,33	85,03
Random Over Sampling	**91,66**	91,15	85,82	88,98	90,99
SMOTE	**93,51**	90,35	84,16	87,71	90,82
SMOTE- Tomek	90,41	**91,15**	87,06	88,49	89,99

TABLE VI. SELECTION OF AUC RESULT FOR 5644 RT-PCR DATA

Sampling Technique	Extra Trees	Light Gradient Boosting	AdaBoost	Gradient Boosting	Random Forest
Without imbalance sampling method	0,9395	**0,9568**	0,9281	0,9469	0,9429
Random Under Sampling	0,9634	**0,9693**	0,9501	0,9515	0,9654
Random Over Sampling	0,9395	0,9544	0,9589	**0,9594**	0,9464
SMOTE	0,9474	**0,9544**	0,9415	0,9395	0,937
SMOTE- *Tomek*	0,9479	**0,9544**	0,9415	0,9395	0,9375

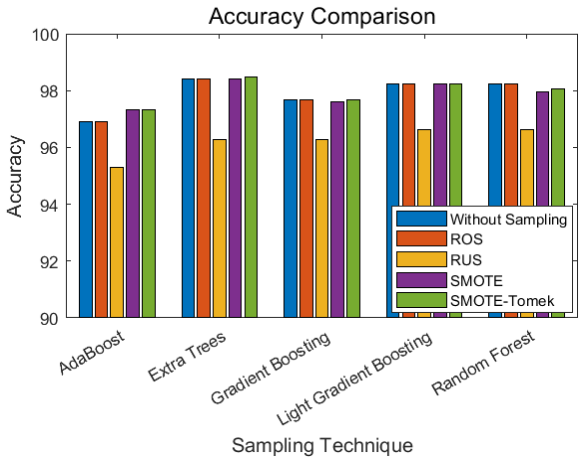


Fig. 2. The Accuracy of the Existing Classification Methods for Covid-19 Prediction.

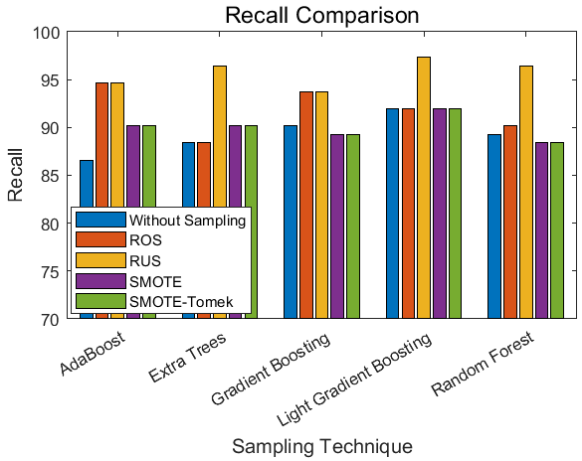


Fig. 3. The Recall of the Existing Classification Methods for Covid-19 Prediction.

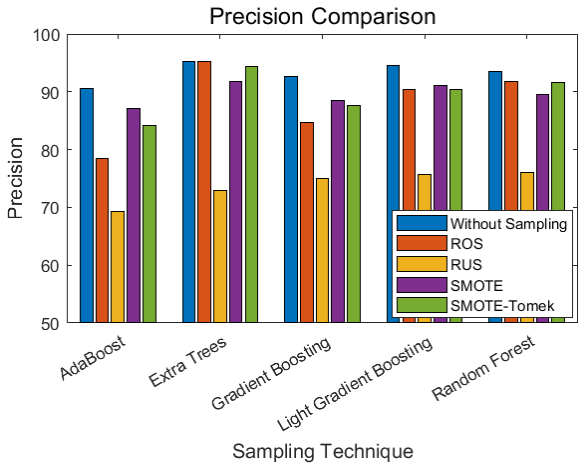


Fig. 4. The Precision of the Existing Classification Methods for Covid-19 Prediction.

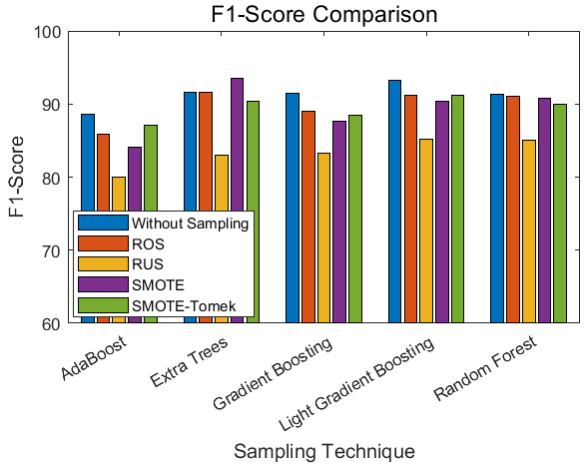


Fig. 5. The F1-score of the Existing Classification Methods for Covid-19 Prediction.

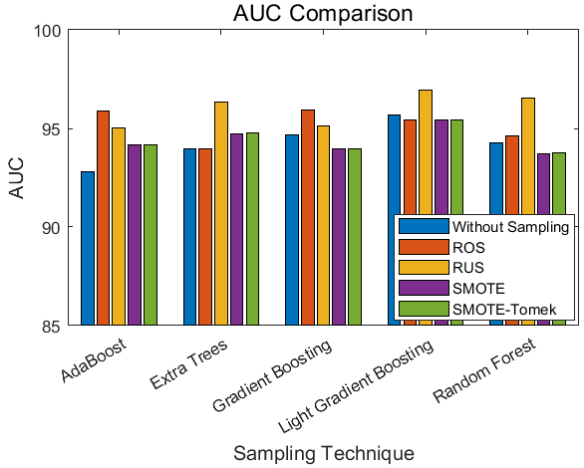


Fig. 6. The AUC of the Existing Classification Methods for Covid-19 Prediction.

The best AUC was produced by light gradient boosting with RUS sampling technique. Light gradient boosting with RUS sampling technique produces AUC score of 0.9693. It can be concluded that the best model that has improved majority of performance in terms of accuracy, precision, recall, f1-score and AUC is light gradient boosting. Light Gradient Boosting produces the best accuracy of 98.49%, recall on the RUS sampling technique is 97.32% and AUC is 0.9693. Furthermore, hyperparamater tuning tests were carried out to optimize the results of Light Gradient Boosting. The parameters used in the Hyperparameter tuning are listed in Table VII.

TABLE VII. SELECTION OF LIGHT GRADIENT BOOSTING PARAMETER ON HYPERPARAMETER TUNING GRID SEARCH

Parameters	**Value**
n_estimators	100, 400, 10
min_child_weight	3, 20, 2
colsample_bytree	0.4, 1.0
max_deph	5, 15, 2
num_leaves	8, 40
min_child_weight	10,30
Learning_rate	0.01,1

After going through the Grid Search process, the best parameters were found that could be tested on the Light Gradient Boosting model. These parameters can be seen in Table VIII.

TABLE VIII. SELECTION OF LIGHT GRADIENT BOOSTING PARAMETERS

Parameters	**Value**	**Parameters**	**Value**
boosting_type	'gbdt',	*n_jobs*	-1,
class_weight	None,	*num_leaves*	40,
colsample_bytree	0.4,	*objective*	None,
importance_type	'split',	*random_state*	None,
learning_rate	0.01,	*reg_alpha*	0.0,
'max_depth	15,	*reg_lambda*	0.0,
min_child_samples	10,	*silent*	True,
min_child_weight	3,	*subsample*	1.0,
min_split_gain	0.0,	*subsample_for_bin*	200000,
n_estimators	400,	*subsample_freq*	0

Table IX is a comparison of light gradient boosting before optimization of hyper parameter tuning and after optimization of hyper parameter tuning.

TABLE IX. SELECTION OF COMPARISON OF LIGHT GRADIENT BOOSTING

Evaluation	***without sampling***	***ROS***	***RUS***	**SMOTE**	**SMOTE-*Tomek***
Accuracy	97.78	98.58	97.25	98.31	98.31
Recall	97.78	98.58	97.25	98.31	98.31
Precision	97.83	98.61	97.65	98.34	98.34
F1-Score	97.83	98.61	97.65	98.34	98.34
AUC	95.68	96.82	96.88	95.88	95.88

The hyper parameter tuning has increased the accuracy of light gradient boosting with an accuracy of 98.58%. The comparison of recall light gradient boosting has increased in almost all tests using sampling techniques. Random forest before the sampling technique was 92.59%. The comparison of F1-score light gradient boosting after hyperparamerer tuning achieved 98.61% on the ROS sampling technique. Based on the results, it can be concluded that light gradient boosting with hyperameter tuning can improve the accuracy, recall, precision, F1-score and AUC. The use of the ROS sampling technique has some advantages in terms of accuracy, recall, precision, f1-score. With the conclusion that the results are 98.58% accuracy, 98.58% recall, 98.61% precision, f1-Score 98.61% and AUC 0.9682%. Based on the results obtained, the results of feature importance are shown in Fig. 7.

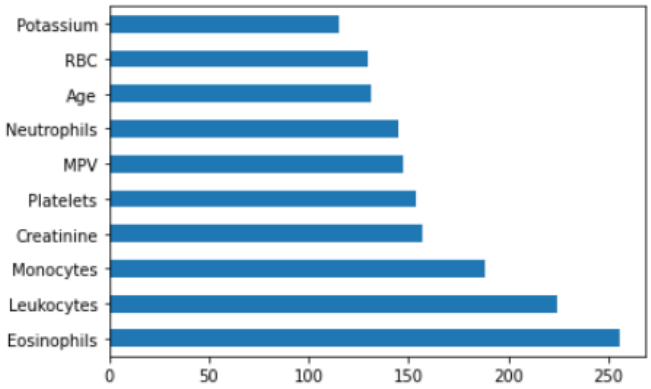


Fig. 7. Importance Features.

Based on Fig. 7, it shows that the first order important features in eosinophiles are. Followed by leukocytes, monocytes, creatinine, platelets, MPV, neutrophils, age, RBC and potassium. The addition of age in the proposed test becomes the seventh most important feature of the best model. The comparison with related research was conducted to assess the performance of the proposed research, the comparison results is listed in Table X.

TABLE X. SELECTION OF COMPARISON WITH RELATED RESEARCH

Model	Accuracy	Recall	Precision	F1 Score	AUC
proposed method	98,58	98,58	98,61	98,61	96,82
Alves *et al* [13]	88	66	91	76	86
Barbosa *et al* [7]	95.16	96.80	93.80	-	-

Based on the table above, it shows that the proposed model produces the best results for all evaluation matrices compare to the previous related studies. With the results of accuracy 98.58%, recall 98.58%, precision 98.61%, F1-Score 98.61% and AUC 0.9682. The visual comparison with related research studies is shown in Fig. 8.

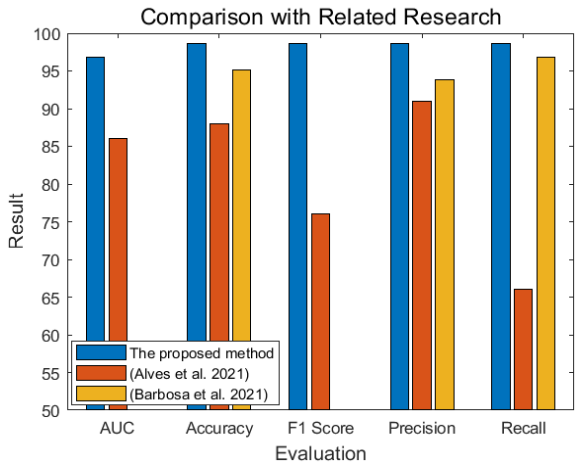


Fig. 8. Comparison with the Existing Studies.

VI. CONCLUSION

This paper has presented various classification methods for COVID-19 prediction. The classification method of light gradient boosting with hyper parameter tuning using ROS sampling technique perform better than the existing the classification methods such as extra trees, random forest, adaboost and gradient boosting for predicting the COVID-19 data. Eosinophils, blood and age parameters has potential become important parameters for COVID-19 prediction. The data was taken from kaggle.com with 5644 data, it shows a classification improvement based on the majority of performance in terms of recall, precision, f1-score and AUC score due to eosinophils, blood and age parameters. Hyper parameter tuning using ROS sampling technique achieved an accuracy of 98.58%, recall of 98.58%, precision of 98.61%, f1-score of 98.61% and AUC of 0.9682. The first important feature in these experiments is eosinophils; it can significantly influence the classification results, while age feature is in the seventh order of important features. In the future research, the proposed model has potential to predict monkey pox disease by identifying important features.

ACKNOWLEDGMENT

This work was supported by Universiti Malaysia Pahang, Universitas Bina Sarana Informatika and Jazan University.

CONFLICT OF INTEREST

On behalf of all authors, the corresponding author states that there is no conflict of interest.

REFERENCES

[1] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses," J. Adv. Res., vol. 24, pp. 91–98, 2020, doi: 10.1016/j.jare.2020.03.005.

[2] H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," J. Autoimmun., vol. 109, no. February, p. 102433, 2020, doi: 10.1016/j.jaut.2020.102433.

[3] Q. Li et al., "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia," N. Engl. J. Med., vol. 382, no. 13, pp. 1199–1207, 2020, doi: 10.1056/nejmoa2001316.

[4] Z. Zhang et al., "Insight into the practical performance of RT-PCR testing for SARS-CoV-2 using serological data: a cohort study," The Lancet Microbe, vol. 2, no. 2, pp. e79–e87, 2021, doi: 10.1016/S2666-5247(20)30200-7.

[5] T. Ai et al., "Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases," Radiology, vol. 296, no. 2, pp. E32–E40, 2020, doi: 10.1148/radiol.2020200642.

[6] E. F. Strasser et al., "Validation of a SARS-CoV-2 RNA RT-PCR assay for high-throughput testing in blood of COVID-19 convalescent plasma donors and patients," Transfusion, vol. 61, no. 2, pp. 368–374, 2021, doi: 10.1111/trf.16178.

[7] V. A. de F. Barbosa et al., "Heg.IA: an intelligent system to support diagnosis of Covid-19 based on blood tests," Res. Biomed. Eng., no. December 2019, 2021, doi: 10.1007/s42600-020-00112-5.

[8] A. 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.

[9] D. Ferrari, A. Motta, M. Strollo, G. Banfi, and M. Locatelli, "Routine blood tests as a potential diagnostic tool for COVID-19," Clin. Chem. Lab. Med., vol. 58, no. 7, pp. 1095–1099, 2020, doi: 10.1515/cclm-2020-0398.

[10] C. M. Goldstein E, Lipsitch M, "On the effect of age on the transmission of SARS-CoV-2 in households, schools and the community," J. Infect. Dis., 2020, doi: https://doi.org/10.1101/2020.07.19.20157362.

[11] M. Dorn et al., "Comparison of machine learning techniques to handle imbalanced COVID-19 CBC datasets," PeerJ Comput. Sci., vol. 7, p. e670, 2021, doi: 10.7717/peerj-cs.670.

[12] B. E. Fan et al., "Hematologic parameters in patients with COVID-19 infection," Am. J. Hematol., vol. 95, no. 6, pp. E131–E134, 2020, doi: 10.1002/ajh.25774.

[13] M. A. Alves et al., "Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs," Comput. Biol. Med., vol. 132, no. March, 2021, doi: 10.1016/j.compbiomed.2021.104335.

[14] A. Banerjee et al., "Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population," Int. Immunopharmacol., vol. 86, no. June, p. 106705, 2020, doi: 10.1016/j.intimp.2020.106705.

[15] E. C. Gök and M. O. Olgun, "SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples," Neural Comput. Appl., vol. 0123456789, 2021, doi: 10.1007/s00521-021-06189-y.

[16] J. Large, J. Lines, and A. Bagnall, "A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates," Data Min. Knowl. Discov., vol. 33, no. 6, pp. 1674–1709, 2019, doi: 10.1007/s10618-019-00638-y.

[17] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement," Inf., vol. 11, no. 6, 2020, doi: 10.3390/info11060332.

[18] A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Land Cover Classification Using Extremely Randomized Trees: A Kernel Perspective," IEEE Geosci. Remote Sens. Lett., vol. 17, no. 10, pp. 1702–1706, 2020, doi: 10.1109/LGRS.2019.2953778.

[19] K. Nugroho et al., "Improving random forest method to detect hatespeech and offensive word," 2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019, pp. 514–518, 2019, doi: 10.1109/ICOIACT46704.2019.8938451.

[20] S. B. Koduri, L. Gunisetti, C. R. Ramesh, K. Mutyalu, and D. Ganesh, "Prediction of crop production using adaboost regression method Prediction of crop production using adaboost regression method," J. Phys. Conf. Ser., 2019, doi: 10.1088/1742-6596/1228/1/012005.

[21] H. Rao et al., "Feature selection based on artificial bee colony and gradient boosting decision tree," Appl. Soft Comput. J., 2019, doi: 10.1016/j.asoc.2018.10.036.

[22] Y. Ju, G. Sun, Q. Chen, M. Zhang, H. Zhu, and M. U. Rehman, "A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting," IEEE Access, vol. 7, no. c, pp. 28309–28318, 2019, doi: 10.1109/ACCESS.2019.2901920.

[23] Y. Su, "Prediction of air quality based on Gradient Boosting Machine Method," Proc. - 2020 Int. Conf. Big Data Informatiz. Educ. ICBDIE 2020, pp. 395–397, 2020, doi: 10.1109/ICBDIE50010.2020.00099.

[24] H. Khafajeh, "An efficient intrusion detection approach using light gradient boosting," J. Theor. Appl. Inf. Technol., vol. 98, no. 5, pp. 825–835, 2020.

[25] Y. Guo, Y. Zhou, X. Hu, and W. Cheng, "Research on recommendation of insurance products based on random forest," Proc. - 2019 Int. Conf. Mach. Learn. Big Data Bus. Intell. MLBDBI 2019, pp. 308–311, 2019, doi: 10.1109/MLBDBI48998.2019.00069.

[26] M. A. M. Hasan, M. Nasser, S. Ahmad, and K. I. Molla, "Feature Selection for Intrusion Detection Using Random Forest," J. Inf. Secur., vol. 07, no. 03, pp. 129–140, 2016, doi: 10.4236/jis.2016.73009.

[27] E. Rendón, R. Alejo, C. Castorena, F. J. Isidro-Ortega, and E. E. Granda-Gutiérrez, "Data sampling methods to dealwith the big data multi-class imbalance problem," Appl. Sci., vol. 10, no. 4, 2020, doi: 10.3390/app10041276.

[28] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," 2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020, no. May, pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.

[29] X. Li and Q. Zhou, "Research on improving SMOTE algorithms for unbalanced data set classification," Proc. - 2019 Int. Conf. Electron. Eng. Informatics, EEI 2019, pp. 476–480, 2019, doi: 10.1109/EEI48997.2019.00109.

[30] "No Hospital Israelita Albert Einstein, Diagnosis of Covid-19 and its Clinical Spectrum - Ai and Data Science Supporting Clinical Decisions (From 28th Mar to 3st Apr).," Accessed on 15/09/2021., [Online]. Available: https://www.kaggle.com/einsteindata4u/covid19?select=dataset.xlsx (Ac.

[31] M. AlJame, I. Ahmad, A. Imtiaz, and A. Mohammed, "Ensemble learning model for diagnosing COVID-19 from routine blood tests," Informatics Med. Unlocked, vol. 21, p. 100449, 2020, doi: 10.1016/j.imu.2020.100449.

[32] H. A. Prihanditya, "The Implementation of Z-Score Normalization and Boosting Techniques to Increase Accuracy of C4.5 Algorithm in Diagnosing Chronic Kidney Disease," vol. 5, no. 1, pp. 63–69, 2020.

[33] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[34] I. H. Witten and E. Frank, Practical Machine Learning Tools and Techniques, Second. San Francisco: Diane Cerra, 2005.

[35] G. Haixiang, L. Yijing, L. Yanan, L. Xiao, and L. Jinling, "BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification," Eng. Appl. Artif. Intell., vol. 49, pp. 176–193, 2016, doi: 10.1016/j.engappai.2015.09.011.

[36] M. R. Camana Acosta, S. Ahmed, C. E. Garcia, and I. Koo, "Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks," IEEE Access, vol. 8, no. Ml, pp. 19921–19933, 2020, doi: 10.1109/ACCESS.2020.2968934.

[37] L. J. Halawa, A. Wibowo, and F. Ernawan, "Face Recognition Using Faster R-CNN with Inception-V2 Architecture for CCTV Camera," ICICOS 2019 - 3rd Int. Conf. Informatics Comput. Sci. Accel. Informatics Comput. Res. Smarter Soc. Era Ind. 4.0, Proc., Oct. 2019, doi: 10.1109/ICICOS48119.2019.8982383.

[38] M. S. Bin Othman Mustafa, M. Nomani Kabir, F. Ernawan, and W. Jing, "An Enhanced Model for Increasing Awareness of Vocational Students Against Phishing Attacks," 2019 IEEE Int. Conf. Autom. Control Intell. Syst. I2CACIS 2019 - Proc., pp. 10–14, Jun. 2019, doi: 10.1109/I2CACIS.2019.8825070.

[39] I. Khandokar, M. Hasan, F. Ernawan, S. Islam, and M. N. Kabir, "Handwritten character recognition using convolutional neural network," J. Phys. Conf. Ser., vol. 1918, no. 4, Jun. 2021, doi: 10.1088/1742-6596/1918/4/042152.

[40] M.A.T. Mohammed, A.S. Sadiq, R.A. Arshah, F. Ernawan, and S. Mirjalili, “Soft set decision/forecasting system based on hybrid parameter reduction algorithm,” Journal of Telecommunication, Electronic and Computer Engineering, vol. 9, no. 2-7, pp. 143-148, 2017.

[41] A. P. Puspaningrum et al., "Waste Classification Using Support Vector Machine with SIFT-PCA Feature Extraction," 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), 2020, pp. 1-6, doi: 10.1109/ICICoS51170.2020.9298982.

Surface Electromyography Signal Classification for the Detection of Temporomandibular Joint Disorder using Spectral Mapping Method

Dr. Bormane D. S[1]
Department of Electronics and Telecommunication
AISSMS College of Engineering
Pune, India

Roopa B. Kakkeri[2]
Department of Electronics and Telecommunication
AISSMS Institute of Information Technology, Pune, India

R. B. Kakkeri[3]
Department of Electronics and Telecommunication
Sinhgad Academy of Engineering
Pune, India

***Abstract*—Temporomandibular joint Disorder (TMD) is with multifaceted and complex signs and symptoms which makes day to day activities of an individual uneasy. Electromyographic (EMG) processing of related muscles recordings could provide an early and immediate detection of TMD. To detect the TMD using surface electromyography (SEMG) of Masseter and Temporalis muscle with discrete wavelet transform (DWT) using spectral coding. To analyze the data, a new feature selection approach in the spectral domain is proposed. For statistical analyses, SPSS version 24 is employed. The results of the study revealed that the proposed approach was able to improve the accuracy of the classification by implementing a combination of DWT and the Support Vector Machine (SVM). The proposed method also exhibited a significant improvement in its performance in terms of its accuracy with 93%. In addition, the statistical analysis revealed that the model was able to improve the mean rank of the experimental and control group.**

***Keywords*—Temporomandibular joint (TMJ); temporomandibular joint disorder (TMD); surface electromyography (sEMG); spectral mapping; discrete wavelet transform (DWT)**

I. Introduction

TMD is an orofacial disorder, which is characterized by joint noises and limitations in the range of motion. Complex and multidisciplinary symptoms related to the Temporomandibular joints (TMJ) are known to affect different parts of the body. Understanding the pathogenesis and possible reasons of these conditions can be challenging. The most common methods of diagnosing and management of TMD is with clinical evaluation followed by radiographic evaluation of images from, X ray, Magnetic resonance Imaging (MRI), cone beam computed tomography (CBCT) and Orthopantomogram (OPG). sEMG is a complementary tool that can be used to evaluate the efficiency and function of muscles by directly analyzing their electrical potentials. This technique has been widely used to diagnose and monitor the TMD as it is a noninvasive technique [1]. In recent development, EMG were processed using spectral coding where multi resolution signals are decomposed and entropy features are used in decision making of bruxism [2]. Various approaches of detection of TMD were developed in the recent research. The significance of usage of sEMG for the reliable detection of TMD is presented in this paper. For the analysis of a superimposed motor unit action a sEMG application for muscle function and its detection efficiency is outlined in [3]. In [4] the signal represents a weighted sum of various temporal and spatial motions of different electrical muscle activities. The analysis is developed as a set of different muscular movement of variation with course of observing time. A muscle imbalance detection based on muscle activity monitoring for TMD is presented in [5, 6]. The outlined approach in [7, 8] aims in developing the effect of temporary splint usage on the masticatory muscle using surface electromyography. The analysis of masticatory muscle activity after an orthodontic treatment using sEMG is discussed. In observing the deviation of masticatory muscle usage for TMD diagnosis an analysis for varying age group is presented in [9]. EMG signals are used in processing masseter and temporalis muscle using duty factors for different age group, and genders of patients having pain and pain free TMD is outlined in [10]. The analysis of TMD a short term observation of transcutaneous electrical nerve stimulation (TENS) observing pain concentration, pressure pain threshold (PPT) and EMG is outlined in [11]. The analysis of sEMG on the behavior of neck, trunk and masticatory muscle for different groups under rest and maximal voluntary clenching (MVC). Variation to sEMG activity for myofacial pain and non-pain condition is presented in [13, 14]. The variation in sEMG for Temporomandibular joint hypermobility (TMJH) for healthy and effective with mild and severe TMJH is presented in [15, 16, 17]. A discrete wavelet transform approach is proposed to analyze sEMG using Auto regression and the Shannon Entropies [18]. The proposed method with new energy-based spectral coding has advantages over the conventional approach and can provide better diagnosis.

The paper is divided into four sections. The first section introduces the concepts of TMD and EMG processing. The second section outlines the materials and methods. Third section describes the proposed approach for detecting disorder with results and discussion, and at the end section four concludes the study.

II. Materials and Methods

A. SEMG Signal Acquisition

This Experimental Analysis was conducted on 100 individuals with an age between 18 to 60 years. The subjects were notified about the procedure and a written informed consent is taken in accordance with the declaration of Helsinki [6]. Many people were omitted from the research because they had a history of orthodontic treatment, lip negligence, or prior tooth restoration [19]. The subjects underwent a complete dental history followed by radiographic evaluation using an Orthopantomography (OPG) by the Dental Practitioner. Based on the Doctors assessment, subjects were divided into two groups. The first group was healthy subjects called control and the second group was subjects with TMD called experimental group. Masseter and Temporalis Muscles of both groups were analyzed by recording the sEMG . The subjects were seated in a comfortable position before the recording was carried out. Since their jaw muscles reacted to changes in head position, no movement was allowed. They were next shown how to hold the mandible relaxed and intercuspal. They were also shown how to utilize cotton rolls to aid in clenching. Signals from the surface of the muscles were collected using a 2-channel electromyography machine. At the beginning of the sEMG recording, volunteer's skin is cleaned with 70% alcohol before the electrodes are placed on it. The participants were monitored for 10 seconds during which they were subjected to rest and clenching activities of 5 seconds each. Both scenarios were recorded using simultaneous electromyographic signals. The recordings were then processed using a 12-bit A/D converter. They are also subjected to a cut-off frequency of 10Hz to 1500Hz.After recording, the signal is filtered through a digital filter with a pass-band of 10Hz to 500Hz. The data were then analyzed using a computer algorithm known as MATLAB 2019. Processing of sEMG is carried out using time frequency analysis with wavelets.

B. SEMG Signal Processing

The proposed system architecture for EMG processing in TMD detection is shown in Fig. 1. The process involves in successive filtration of signal using high pass and low pass filters for a given scale levels. The wavelet function can be performed using two different algorithms: the Haar wavelet algorithm and the Daubechies algorithm (3). The goal of the wavelet function is to match the signal to the wavelet function that's being processed.

In this study Daubechies-4 wavelets with five levels of scaling is applied after comparing and getting better results with other wavelets and the decomposition is developed as a separation of finer spectral bands and residual elements given by

$$B[t] = \sum_{t=1}^{n} \xi[t] + \gamma[t] \quad (1)$$

Where, B[t] represent the spectral bands for a given scale level of 't' processed over a period of t=1..n, which is observed to be sum of spectral bands ξ and residual decomposed band γ. The spectral bands are derived from a set of filter blocks which are cascaded to develop a hierarchical decomposition of given signal, which results in filter spectral bands and residual components.

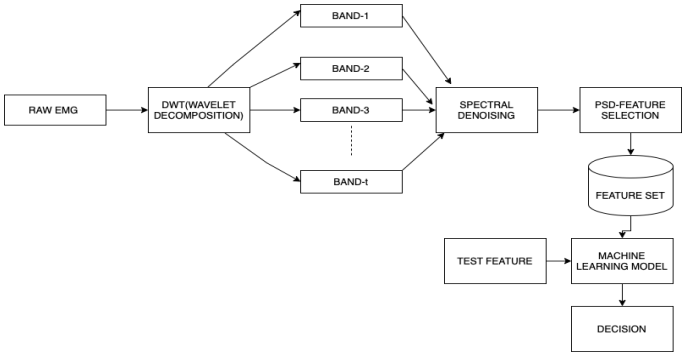


Fig. 1. Framework for Proposed Approach of TMD Detection: Raw EMG is Decomposed using DWT with Multiresolution Approach using Daubichew-4 Filter with Five Levels of Scaling for Denoising and the Reconstructed Signal is extracted with Features and Selected Feature.

The decomposed bands for the EMG signal represent a set of sub-spectral bands $\{\chi_0, \chi_1, \chi_3 \dots \dots . \chi_n\}$ which are derived by the signal filtrations of successive residual component of previous scale. The decomposed bands are defined as a set υ which are the subset of processing signal S given by,

$$\upsilon = \{\chi_0, \chi_1, \chi_3 \dots \dots . \chi_n\} \in S \quad (2)$$

In filtration, the maximum spectral band of each of the decomposed band is computed and average value is considered as filtering coefficients defined as,

$$f_i = \max(\chi_i)_{i=1 \text{ to } t} \quad (3)$$

$$F = \{f_1, f_2, f_3 \dots . f_t\} \quad (4)$$

$$F_{coeff} = \text{avg}(F) \quad (5)$$

The average over all decomposed band defines the distribution of varying peaks for each band. Maximum peaks of each band reflect the maximum spectral limit of the processing band. The filtration process performs a convolution operation of the decomposed band using computed filter coefficients for distortion removal. The filtration is performed as,

$$D_i = \text{conv}(\chi_i, F_{coeff}) \quad (6)$$

The denoised signal (Di) is processed for feature selection where the spectral densities of the filtered bands are computed. Spectral density illustrates the spectral energy concentric on a band. The spectral density is defined as the power spectral density (PSD) given as,

$$\text{PSD } (P_i) = \frac{1}{n}(\sum_{i=1}^{n} D_I) \quad (7)$$

Where, 'n' is the total number of coefficients in the band. For 't' scale bands the PSD are a set of PSD's given as {P_1, P_2, P_3 …… P_t}. In selecting features the computed spectral density is correlated and bands with PSD above the limiting value is considered as feature. The selection of feature vector is outlined in below algorithm.

Proposed Algorithm for Feature Selection:

Input: Denoised spectral bands of EMG signal (Di)

Output: Selected features

Process

Compute PSD (P_i) for obtained denoised band Di,

$$PSD\ (P_i) = \frac{1}{n}\left(\sum_{i=1}^{n} D_i\right)$$

Compute limiting threshold (L) given as,

$$= \frac{1}{4}\max(P_i)$$

Select feature using correlative method defined as,

$$Ft\ = \begin{cases} P_{i,}\ if\ P_i > L, \\ \\ 0\ else \end{cases}$$

End process

Features were selected for a limiting threshold value of 25%. Because most of the dominant values will be above this threshold. Lower values may be considered but might contain noise. The selected feature sets are trained for sparsely disturbed EMG signals of different test cases and a training process is performed to create a learning feature table. The learning feature sets is used for training a support vector machine (SVM) and testing is performed using the multi class model of SVM classifier. The processing system operates for training and testing operation, where a training process is an offline process and testing is developed as an online process. The simulation result and observations for the proposed system is presented in following section.

III. Results

A. Results of Signal Processing and Classification Model

The signal processing window of 300 coefficients is considered in processing the EMG signal. The processing window signals is iteratively processed to minimize the processing overhead and reflects a finer multi resolution details for processing signal. The EMG signal is processed for distortions with additive Gaussian noise, which are observed in the signaling domain generated due to surrounding environment and processing units. Multi resolution spectral bands using wavelet transformation is shown in Fig. 2. The decomposed bands reflect the variation of signal under different frequency resolutions.

The decomposed bands illustrate a magnitude variation under different frequency resolution [16]. These finer details give more selective characteristic of processing signal. The spectral energy density of each isolated band computed as PSD. A test of 10 iterations with varying Variance value from 0 to 1 is performed. The denoising efficiency is measured by an average Means square error (MSE), Peak signal to noise ratio (PSNR) and Root mean square error (RMSE) values.

The MSE is given by,

$$MSE = \frac{\sum_{i=1}^{n}(x-x')^2}{n} \quad (8)$$

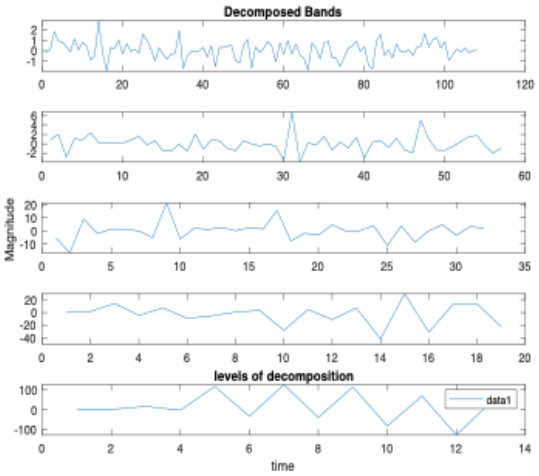


Fig. 2. Decomposed Spectral Bands for the Surface EMG Signal for Level 5 and Db4 Wavelet.

Where x is the actual signal of processing coefficients and x^' is the de-noised signal coefficients.

PSNR is given by,

$$PSNR = log_{10}\frac{peak(x')}{peak(x)} \quad (9)$$

and RMSE is given by,

$$RMSE = \surd MSE \quad (10)$$

Observations for the developed approach spectral mapped approach (SMAP) is compared with the existing approach of soft thresholding and auto regression model (AR) [18] for denoising performance. The observation of the developed approaches for denoising for different test samples is presented in Table I.

TABLE I. Denoising for Different Test Sample

Test sample	**Method**	**PSNR (dB)**	**RMSE**	**Time (Sec)**
S1	Soft threshold	34.9088	2.5115	0.156
	AR	45.4041	1.1643	0.031
	SMAP	50.8703	1.1756	0.015
S2	Soft threshold	34.4641	2.5064	0.175
	AR	45.4121	1.1642	0.029
	SMAP	50.8703	1.1756	0.019

The spectral coding of signal denoising eliminate distortion using a period of observation, wherein a discrete observation generate filtration for observing coefficients only which has lower filtration performance. More effective denoising results into an accurate signal representation. This result to improve the accuracy of detection.

The peak signal to noise ratio (PSNR) [18] value defines the signal strength to distortion in the processing signal. A higher PSNR defines higher signal strength in retrieved signal in reference to distortion level. The proposed method attained 51dB of PSNR which is 20dB and 8dB higher than the existing soft threshold and auto regression model respectively. Root

mean square error (RMSE) defines the standard deviation of the predicted signal compared to the original signal. The RMSE of the proposed method is 1.9 times lower than auto regression model and 2.1 time lower compared to soft threshold method.

The spectral peak values of the denoised signals are considered as a relative feature for EMG signal analysis. Detected peak levels of the processing signal is shown in Fig. 3. The spectral features are used for training the SVM model and classification of the test signal is performed for the developed multi class model [18]. Testing of developed approach is made over a sparse dataset of captured EMG samples under various postures. A k-fold test is performed where 1/kth part of the database is used for training and remaining is used for testing. The approach presented is outlined in MATLAB tool and validated for the retrieval Accuracy, sensitivity, specificity, Recall Rate, precision and computation time parameter [6]. Processing sEMG signal is captured at 5mv/sec for a period of 2 msec period.

The validation of the developed approach is performed using retrieval Accuracy, sensitivity, specificity, Recall Rate, precision and computation time parameter.

The observation for the developed system is computed for different test cases measuring the evaluating parameters. The observation for the developed approach for different classifier models, k-fold tests and varying interference levels are presented. The developed approach of Spectral mapped based classification is compared with the existing approach of Quadrant discriminant analysis and Naïve Bayes method [19]. Observation for different K-fold test for K=2 and 3 is given in Table II.

The classification performance of the proposed approach using spectral mapping method has 93% of accuracy with increases of noise variance the accuracy marginally reduced due to spectral domain processing compared to existing Quadrant discriminant analysis, KNN and Naïve Bayes classifier [19].

B. Statistical Data Analysis

Data was collected from a total of 100 respondents, who were divided into two groups: control (healthy) and experimental (patients). Of the 50 respondents in the control group, 29 were female and 21 were male. In the experimental group, there are 26 females and 24 males out of a total of 50 responses. Descriptive statistics were used to analyse the various variables within the group and sex. Two-way factorial analyses were performed to compare the mean values. The mean maximum amplitude of the subjects during rest and MVC was also estimated. The features were selected for the groups of female and male.

For all of the features, descriptive statistics were first calculated. The kolmogorov Smirnov test is used to check for normalcy, followed by a retest using the nomal Q- Q plot. We discovered that the data is not normally distributed and that the Kolmogorov-Smirnov P-value (.000) is less than.050. Furthermore, the largest observation is away from the primary diagonal in the Q-Q graph, indicating that the data is not regularly distributed. The data was then corrected for normality using log transformation, reciprocal transformation, and square root transformation. However, this method failed to convert the data into normal distribution. The results are presented in Table III. The non-parametric Mann-Whitney U test was used to investigate the significant difference between left temporalis and right temporalis rest and MVC among the participants.

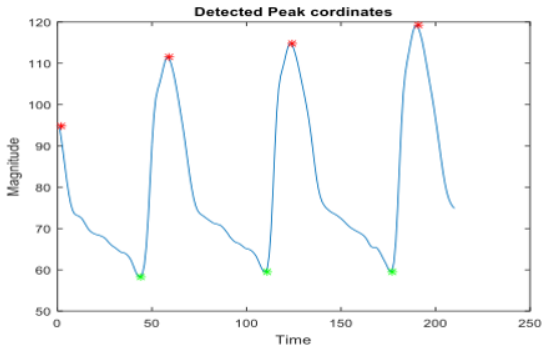


Fig. 3. Peak Detected Energy Coefficients which are the Proposed Features for the Classification.

TABLE II. DENOISING METHODS FOR DEVELOPED SPECTRAL MAPPED APPROACH (SMAP)

K-fold	Method	Accuracy	Sensitivity	Specificity	Precision	Processing Time (sec)
K=2	Quadrant discriminant analysis	41.09	0.85	0.12	0.38	2.03
	KNN	79.07	0.89	0.20	0.89	0.69
	Naïve Bayes	88.09	0.94	0.233	0.92	0.17
	Spectral mapped approach	92.9	0.96	0.253	0.94	0.041
K=3	Quadrant discriminant analysis	43.87	0.86	0.09	0.43	2.05
	KNN	80.03	0.89	0.21	0.87	0.70
	Naïve Bayes	87.39	0.94	0.238	0.91	0.17
	Spectral mapped approach	92.63	0.96	0.251	0.93	0.043

TABLE III. MANN-WHITNEY TEST STATISTIC

Group		Right Temporalis Rest	Left Temporalis Rest	Right Temporalis MVC	Left Temporalis MVC
Control	Mann-Whitney U	51.000	53.000	23.000	77.500
	Wilcoxon W	486.000	488.000	458.000	512.500
	Z	-5.500	-5.464	-5.970	-5.050
	Asymp. Sig. (2-tailed)	.000	.000	.000	.000
Experimental	Mann-Whitney U	15.500	25.500	19.000	116.000
	Wilcoxon W	246.500	256.500	250.000	347.000
	Z	-5.381	-5.155	-5.302	-3.095
	Asymp. Sig. (2-tailed)	.000	.000	.000	.002

IV. DISCUSSION

In this study, we evaluated the effectiveness of surface electromyography in detecting the temporomandibular disorders (TMD) with the help of time-frequency domain analysis using discrete wavelet transform. The novel approach for this study is to take out the spectral peak features after denoising the EMG signal for further classification and detection. The feature representation for diagnosis using entropy based approach outlined in [18] is constraint with the magnitude variation. The proposed approach offers an integral advantage of filtration and feature selection as single processing unit. Discrete wavelet transformation has shown a significant advantage in frequency domain decomposition because of the property multi resolution coding.

The study was limited due to the small number of subjects and the heterogeneous group. Since this disorder has complex signs and symptoms, we have considered only masseter and temporalis muscle activity. Other than muscle activity, Temporomandibular Joint sounds could be analyzed. Although the results of the study indicated that surface electromyography can identify patients with TMD, Surface electromyography is noninvasive and easier to operate but during recording possibility of noises due to surrounding or sensors is accumulated which needs to be preprocessed before the analysis. Intramuscular electromyogram (EMG) may help in pattern recognition. However, the individual will experience discomfort as a result of the intrusive electrodes. The goal of the study was to determine if a complex disease can be represented through a phenotype that allows for a straightforward clinical assessment. It also wanted to know how much of the variation in the development of the neuromuscular system can be reduced through the use of morphologies.

V. CONCLUSION

The purpose of this study was to devise a new method for decomposing and classifying surface EMG signals using spectral coding approach in order to extract useful information. The proposed method is based on spectral domain signal denoising, which highlights the lowest distortion and allows the system to retrieve the smallest signal feasible. The resulting technology can greatly enhance signal retrieval accuracy. Spectral energy peaks as feature sets when applied to multiclass machine learning models performed better with accuracy and other parameters. Support Vector Machine enhanced performance and it can help with the accurate diagnosis. The subjects with TMD exhibited significant different muscle activity at than those with the control group. The theme of the study also highlighted the importance of having a comprehensive understanding of the complex disease. This is because the lack of a descriptive taxonomy can hinder the development of effective treatments. Our understanding of TMJ issues lags behind that of pain disorders. The various themes of the study also highlighted the importance of having a comprehensive understanding of the complex disease. This is because the lack of a descriptive taxonomy can hinder the development of effective treatments. The multiple approaches that are currently utilized in the research and development of TMD should be combined with a systematic strategy.

REFERENCES

[1] Ishii T, Narita N, Endo H. Evaluation of jaw and neck muscle activities while chewing using EMG-EMG transfer function and EMG-EMG coherence function analyses in healthy subjects. Physiol Behav [Internet]. 2016;160:35–42. Available from: http://dx.doi.org/10.1016/j.physbeh.2016.03.023.

[2] K S DP, Rathika rai D, Easwaran DM, Easwaran DB. Temperomandibular disorders Clinical and Modern Method In Differential Diagnosis. IOSR J Dent Med Sci. 2014;13(9):01–7.

[3] Altın C, Er O. Comparison of Different Time and Frequency Domain Feature Extraction Methods on Elbow Gesture's EMG. Eur J Interdiscip Stud. 2016;5(1):35.

[4] Gokgoz E, Subasi A. Comparison of decision tree algorithms for EMG signal classification using DWT. Biomed Signal Process Control [Internet]. 2015;18:138–44. Available from: http://dx.doi.org/10.1016/j.bspc.2014.12.005.

[5] Bormane DDS, Kakkeri RB. Detection of Temporomandibular Joint Disorder Using Surface Electromypgraphy by Supervised Classification Models. Mater Today Proc [Internet]. 2021;(xxxx). Available from: https://doi.org/10.1016/j.matpr.2021.07.375.

[6] Klasser GD, Greene CS. Oral appliances in the management of temporomandibular disorders. Oral Surgery, Oral Med Oral Pathol Oral Radiol Endodontology [Internet]. 2009;107(2):212–23. Available from: http://dx.doi.org/10.1016/j.tripleo.2008.10.007.

[7] Szyszka-Sommerfeld L, Machoy M, Lipski M, Woźniak K. The Diagnostic Value of Electromyography in Identifying Patients With Pain-Related Temporomandibular Disorders. Front Neurol. 2019;10.

[8] Sapsanis C, Georgoulas G, Tzes A, Lymberopoulos D. Improving EMG based classification of basic hand movements using EMD. Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS. 2013;5754–7.

[9] Akbulut N, Altan A, Akbulut S, Atakan C. Evaluation of the 3 mm Thickness Splint Therapy on Temporomandibular Joint Disorders (TMDs). Pain Res Manag. 2018;2018.

[10] Bianchi J, de Oliveira Ruellas AC, Gonçalves JR, Paniagua B, Prieto JC, Styner M, et al. Osteoarthritis of the Temporomandibular Joint can be diagnosed earlier using biomarkers and machine learning. Sci Rep. 2020;10(1):1–14.

[11] Altın C, Er O. Comparison of Different Time and Frequency Domain Feature Extraction Methods on Elbow Gesture's EMG. Eur J Interdiscip Stud. 2016;5(1):35.

[12] Pitta NC, Nitsch GS, Machado MB, de Oliveira AS. Activation time analysis and electromyographic fatigue in patients with temporomandibular disorders during clenching. J Electromyogr Kinesiol. 2015;25(4).

[13] SANTOS AC dos, SILVA CAB da. Surface Electromyography of Masseter and Temporal Muscles With Use Percentage While Chewing on Candidates for Gastroplasty. ABCD Arq Bras Cir Dig (São Paulo). 2016;29(suppl 1):48–52.

[14] Ferreira AP de L, Da Costa DRA, De Oliveira AIS, Carvalho EAN, Conti PCR, Costa YM, et al. Short-term transcutaneous electrical nerve stimulation reduces pain and improves the masticatory muscle activity in temporomandibular disorder patients: A randomized controlled trial. J Appl Oral Sci. 2017;25(2):112–20.

[15] Ishii T, Narita N, Endo H. Evaluation of jaw and neck muscle activities while chewing using EMG-EMG transfer function and EMG-EMG coherence function analyses in healthy subjects. Physiol Behav [Internet]. 2016;160:35–42. Available from: http://dx.doi.org/10.1016/j.physbeh.2016.03.023.

[16] Karelina AN, Geletin PN, Ginali N V, Romanov AS. Method of Diagnosis of the Temporomandibular Joint Disorders. Sci J Res Dent [Internet]. 2017;1(3):67–070. Available from: www.scireslit.com.

[17] Khan MMR, Arif RB, Siddique AB, Oishe MR. Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI machine learning repository. 4th Int Conf Electr Eng Inf Commun Technol iCEEiCT 2018. 2019;124–9.

[18] Sonmezocak T, Kurt S. Detection of EMG signals by neural networks using autoregression and wavelet entropy for bruxism diagnosis. Elektron ir Elektrotechnika. 2021;27(2):11–21.

[19] Gokgoz E, Subasi A. Comparison of decision tree algorithms for EMG signal classification using DWT. Biomed Signal Process Control [Internet]. 2015;18:138–44. Available from: http://dx.doi.org/10.1016/j.bspc.2014.12.005.

A Deep Learning Approach for Viral DNA Sequence Classification using Genetic Algorithm

Ahmed El-Tohamy, Huda Amin Maghwary, Nagwa Badr
Information Systems Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt

***Abstract*—DNA sequence classification is one of the major challenges in biological data processing. The identification and classification of novel viral genome sequences drastically help in reducing the dangers of a viral outbreak like COVID-19. The more accurate the classification of these viruses, the faster a vaccine can be produced to counter them. Thus, more accurate methods should be utilized to classify the viral DNA. This research proposes a hybrid deep learning model for efficient viral DNA sequence classification. A genetic algorithm (GA) was utilized for weight optimization with Convolutional Neural Networks (CNN) architecture. Furthermore, Long Short-Term Memory (LSTM) as well as Bidirectional CNN-LSTM model architectures are employed. Encoding methods are needed to transform the DNA into numeric format for the proposed model. Three different encoding methods to represent DNA sequences as input to the proposed model were experimented: k-mer, label encoding, and one hot vector encoding. Furthermore, an efficient oversampling method was applied to overcome the imbalanced dataset issues. The performance of the proposed GA optimized CNN hybrid model using label encoding achieved the highest classification accuracy of 94.88% compared with other encoding methods.**

Keywords*—Deep learning; sequence classification; convolutional neural networks; genetic algorithm; sequence encoding

I. Introduction

Viruses are the leading cause of infectious diseases and have a harmful impact on the human population. Recent examples of such viruses include COVID-19, SARS, and MERS. As a result of viral outbreaks, new vaccines have been developed for such pathogens [1]. When it comes to virus subtyping and taxonomy classification, the classification of a virus's genomic sequence is extremely vital to analyze and understand for faster production of such vaccines. A virus's genome is either made up of DNA or RNA, and it is referred to as a DNA virus or an RNA virus, accordingly [2]. An organism's genetic code is encoded in the deoxyribonucleic acid (DNA). Adenine (A), thymine (T), cytosine (C), and guanine (G) are the four nucleotides that the DNA consists of. These are referred to as the DNA nucleotide bases [3]. Each type of nucleotide has a binding to its complementary pair on the opposite strand in the double-stranded DNA. Adenine and cytosine form a pair with thymine and guanine, respectively. Single-stranded or double-stranded RNA are both possible for ribonucleic acid. T is replaced by U in RNA. The improvement of phylogenetic and functional research of viruses may be enhanced by the correct classification of genomic sequences [4,5]. Genomic sequences are classified into different groups based on their qualities, traits, or attributes, and this process is known as genomic sequencing classification. The more information is known about the virus, the closer an efficient vaccine can be developed quickly. Because viruses' genomic sequences may have little in common with those of other viruses, it is difficult to classify them. The genomic sequence can be classified using several different approaches. Machine learning models can be trained using well-understood sequences to predict the profile of unknown sequences [6]. As a new branch of machine learning, deep learning has emerged in the last several years. To represent data at increasingly higher levels of abstraction, these models employ multiple non-linear transformations. These models can deal with complex challenges because of their many hidden layers. Many studies have used machine learning and deep learning algorithms to analyze DNA sequences [6,7]. Manual feature extraction is used in these machine learning models [8]. On the other hand, this can lead to various complications, such as selecting features that do not lead to the optimal solution or missing out on key features. Most significantly, the main key features from the DNA dataset extracted are not clear. Besides, it is difficult to extract these features manually using traditional machine learning algorithms. Therefore, an automatic feature selection approach is proposed to overcome this issue. One of the greatest deep-learning methods for extracting important characteristics from a dataset is convolutional neural networks (CNNs) [9,10]. This study proposes an optimized convolutional neural network architecture for DNA sequence classification using genetic algorithm (GA) optimization layer as well as a long short-term memory (LSTM) layer. LSTM is a kind of recurrent neural network (RNN). It can process entire sequences of data effectively [11]. Besides, A genetic algorithm (GA) is proposed to optimize the deep learning model. GA is a heuristic approach inspired by the process of natural selection that is used in computer science and operations research [12]. It is a subclass of evolutionary algorithms (EA) that includes other metaheuristics. Genetic algorithms are commonly employed to generate solutions to optimization and search problems by utilizing bio-inspired operators such as mutation, crossover, and selection [13,14]. A genetic algorithm optimization layer was implemented to improve the accuracy of the classification model. The introduction of evolutionary algorithms such as genetic algorithms showed to optimize deep neural network weight matrix [15]. Thus, optimizing the weight matrix of the

convolutional neural network can achieve a better classification accuracy. It can also give better classification results for the LSTM models as the CNN layer output is used as an input to them. As a proof of concept, the optimized model was compared with and without the proposed GA optimization layer. The accuracy of the model with the GA optimization layer is shown to be better than the model without it. Moreover, a comparison was held using the same dataset with previously implemented models. The used dataset contains more viral sequences that may dominate the learning process which lead to a false increase in the overall accuracy. Therefore, an improved oversampling approach was applied to overcome the imbalanced dataset issue. The main contributions of this paper include a proposal of a hybrid deep learning model for efficient viral DNA sequence classification and an introduction of an optimization evolutionary algorithm to the proposed classification framework to improve the overall accuracy. In addition, an efficient oversampling approach is applied for handling the imbalanced dataset as well as increasing the dataset class variability. Besides, one-hot encoding is newly experimented on the viral DNA sequence dataset as an encoding method whereas k-mer encoding [16] and label encoding was used before. The paper is organized as follows: in Section II, the related work is reviewed. Section III describes the dataset and the different preprocessing techniques applied on the dataset. Then, the proposed approach is presented. In Section IV, the experimental results and comparisons with other models are demonstrated. Finally, the paper is concluded in Section V.

II. Related Work

Different studies employed several models and techniques for the classification of viral sequences. In [17], a new approach for classifying the Avian Influenza A viral (AIAV) sequences of the hemagglutinin (HA) and neuraminidase (NA) genes into subtypes using DNA sequence data and physicochemical properties is proposed. Mainly using machine learning techniques, four different classifiers, Naïve Bayes, Support Vector Machine (SVM), K-nearest neighbor (KNN), and Decision Tree were compared. The Decision Tree achieved the best accuracy of 95%.

In [18], the author proposed three models for the classification of different viral DNA sequences using raw DNA sequence data. The three classification models were CNN, long short-term memory (LSTM), and convolutional neural network bidirectional long short-term memory (CNN-Bidirectional LSTM). He used the Synthetic Minority Oversampling Technique (SMOTE) algorithm for data oversampling to overcome imbalanced dataset problem with two encoding methods: label encoding and k-mer encoding. Results showed that k-mer encoding achieved the best results with 93.16% accuracy of the CNN model.

In [19], the author used Random Forest and Artificial Neural Network models with metagenomic sequences that were taxonomically sorted into virus and non-virus categories. The algorithms attained accuracy considerably above the level of chance, with an area under the ROC curve of 0.79. There were two codons (TCG and CGC) that showed the most discriminative features for classification.

In [20], the author utilized combining two classification algorithms with ensemble techniques such as Xgboost and random Forest to improve the accuracy of classifying DNA sequence splice junction types for small example datasets. They achieved an accuracy of 96.24% for Xgboost and 95.11% for Random Forest.

The author in [9] developed a novel method for classifying DNA sequences using a convolutional neural network and treating the sequences as text input. The author employed one-hot vectors to represent sequences as input to the model. The approach was evaluated on 12 DNA sequence datasets. Significant improvements were found in all the previous models using his proposed approach for DNA sequence classification with improved accuracy up to 6.12% on the H3K4me3 dataset.

Most of the existing works tend to focus on training the classification models without any kinds of optimization both on the preprocessing step and prior to the classification step. Therefore, in this research, a hybrid deep learning model with a genetic algorithm optimization layer is proposed. The genetic algorithm layer is applied to optimize the weights of the CNN model. The CNN model is then utilized for classification as a separate model as well as an input to the LSTM and CNN-LSTM Bidirectional models. This will greatly improve the overall accuracy. Thus, the classification method uses the optimized genetic algorithm to generate CNN weights. As a prior step in data preprocessing, Adaptive Synthetic Sampling Approach (ADASYN) is used to handle the imbalanced dataset issues.

III. Materials and Methods

A. Dataset

The DNA dataset was extracted from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov). NCBI contains entire DNA sequences for viruses which is publicly available. The acquired virus DNA sequence datasets are COVID, SARS, MERS, dengue, hepatitis, and influenza. In addition, entire DNA sequences for Zika and EBOLA viruses were collected. A FASTA file for each sequence data was collected and downloaded for complete genetic sequences of each class label with sequence ranges from 8 to 38,012 nucleotides. The collected dataset consists of 86,637 inputs. A distribution of each class label and the count of samples in each label is shown in Table I.

TABLE I. Dataset Class Distribution

Class Label	Number of Samples
COVID	45216
SARS	7311
MERS	6735
Dengue	1994
Hepatitis	8577
Influenza	11862
Zika	1920
EBOLA	3022

As shown in Table I, the minority classes like MERS, SARS, Zika, Ebola, and Dengue have low counts unlike COVID, Hepatitis, and Influenza. To overcome this imbalanced dataset issue, the Adaptive Synthetic Sampling Approach (ADASYN) [21] was applied. ADASYN is used to generate synthetic data for the minority classes to oversample them to match the majority classes. ADASYN is a generalized form of the SMOTE (Synthetic Minority Oversampling Technique) algorithm. SMOTE [22] is an oversampling technique in which synthetic samples are generated for the minority data class. Random oversampling can lead to overfitting, which is why this approach helps alleviate that problem [23]. The main difference between ADASYN and SMOTE is that by using ADASYN the number of synthetic instances generated for samples that are difficult to learn is determined by taking the density distribution into account. As a result, difficult-to-learn samples can be used to adaptively alter decision boundaries. ADASYN works by locating the closest k-nearest neighbors of the minority class using Euclidean distance. Then, it chooses a random neighbor, and a line is constructed between the neighbor and the minority class data point. A synthetic sample is generated between them. Fig. 1 demonstrates how the synthetic data points are generated using ADASYN.

B. Data Preproccessing

The most important aspect of both machine learning and deep learning algorithms is preprocessing of data. It affects the accuracy of the proposed model drastically. DNA sequences, unlike text data, are sequences of consecutive letters without a space between them. No words or phrases can be found in the DNA sequence. As a result, k-mer encoding [16] is used for converting DNA sequences into word sequences. This preserves the nucleotide positions of each word in the sequence. Two vector encoding methods, one hot vector encoding, and label encoding are also used to represent the numerical values of the sequences [24]. One hot vector encoding, and label encoding are used because in contrast to image data, which is represented as a two-dimensional numerical matrix as an input to the CNN, text data is represented as a one-dimensional series of consecutive characters. As a result, it must be converted to numerical values to use as the input for the CNN model. A demonstration of both sequence encodings is shown in Fig. 2.

Thus, encoding is the process of transforming nucleotide categorical data into numerical data. In this research paper, three different types of encoding methods, Label encoding, one hot vector encoding, and k-mer encoding, were experimented with separately to encode the DNA sequence and convert it to the suitable numerical form for deep learning. Label encoding is a popular method for representing categorical data as binary vectors efficiently. For each of the four classes of nucleotides (A, T, G, and C), each one is represented as a number to form an array. A is given the value of 1, C is given the value of 2, G is given the value of 3, and finally, T is given the value of 4. An example sequence of (AACG) will be represented as an array of integers of (1,1,2,3). In decimal-binary vector encoding, one-hot vector encoding for DNA sequences is another way of representing nucleotide sequence data in numerical vector representation. Each nucleotide is represented by a binary vector of length 4. A is represented as (1,0,0,0), C as (0,1,0,0), G as (0,0,1,0) and T as (0,0,0,1). Each nucleotide holds a vector representation of 4x1 dimension. Finally, k-mer encoding transforms the complete DNA sequence into smaller substrings of length k, which represents a word. These words can be used effectively in natural language processing techniques.

C. Classification Methods

Three deep learning models were applied. One model consists of the CNN layer only. The other two models consist of two layers. The first layer of both models is the CNN layer. The CNN layer is used as a feature extraction layer. The output of the CNN layer is given as an input to the second layer. The second layer of the first model is CNN-LSTM. The second layer of the second model is CNN-Bidirectional LSTM. One of the main contributions of this work is applying a standard Genetic Algorithm (GA) to optimize each CNN layer in the models. The GA layer is used to optimize the weights in the CNN layer, which in turn improves the accuracy of the classification models [25,26,27]. Each model is trained and tested using three different encoding methods, label encoding, one-hot vector encoding, and finally using k-mer encoding. A summary of the proposed workflow with the models is shown in Fig. 3.

As demonstrated in Fig. 3, after the data preprocessing the GA layer is utilized to optimize the weights of the CNN layer. Then, the three models are used for the classification process.

This section demonstrates the classification methods in detail. In subsection 1, a detailed demonstration of the proposed genetic algorithm optimization layer will be presented. Following that, in subsections 2 and 3 the used classification models will be explained, respectively.

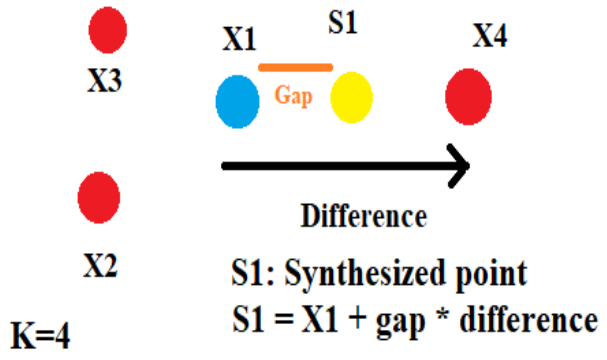


Fig. 1. Generation of Synthetic Data Points using ADASYN with k=4 as an Example and S1 Represents the Synthesized Point of the Minority Class where X_n Represents a Data Point.

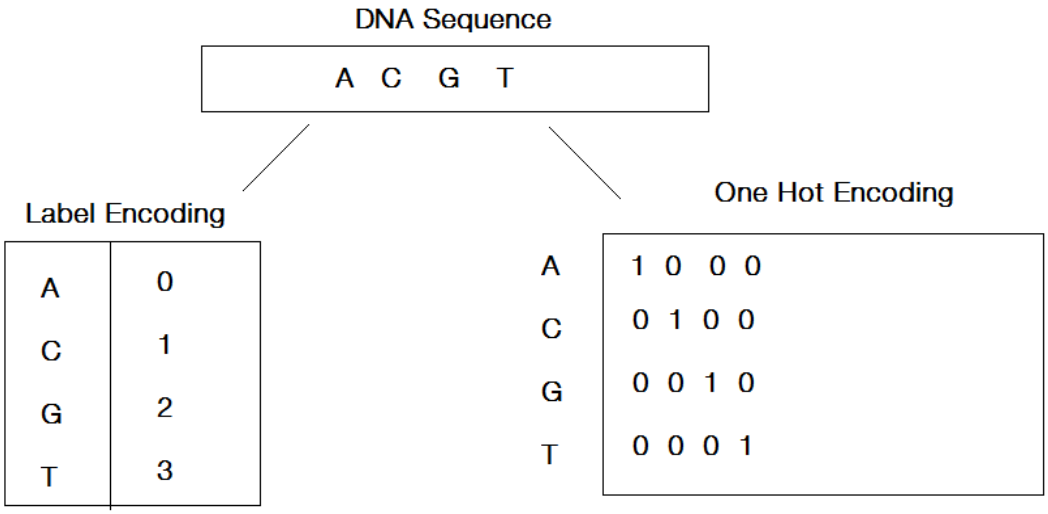


Fig. 2. Difference between One-Hot Encoding and Label Encoding for DNA Sequences.

1) Optimization layer using genetic algorithm (GA): Genetic algorithm [12] relies on biologically inspired operators including mutation, crossover, and selection to produce high-quality solutions to optimization and search problems. GA is mainly a heuristic approach for the optimization of search problems. It is used because the concern is about the optimization of the weights not how much time it takes. Thus, in this research, it is used to optimize the weights of the CNN layer.

The standard GA progression originally proceeds as follows:

- The population's initialization.
- Evaluating each member's fitness.
- Choosing parents to create children for the next generation.
- Parental cross-over to create offspring.
- Randomly mutating the offspring.
- Keep evaluating, reproducing, and mutating until the loss function is optimized.

The following are the proposed steps involved in integrating the GA with the CNN:

- Randomization of initial values of each chromosome.
- Substituting the CNN weights with the values of the selected chromosome.
- Using the newly obtained weights to update the weights of the CNN.
- Calculating the fitness of the present offspring by subtracting the resultant output from the goal output sequence.
- Repeating the simulation for all members of the population.
- Using a roulette strategy for selecting the parents of the next generation.
- Crossover of the parents to produce new offspring.
- Mutating the offspring with a 1% probability of mutation.
- Repeat the previous steps until the evaluation metrics or loss function is optimized. A pseudocode of the proposed GA algorithm is shown in Algorithm 1.

Algorithm 1: Genetic Algorithm for CNN Optimization

Input:
Population Number, *n*
Iterations, *I*

Output: Global best configuration of CNN weights O_{best}

Begin
Generation of population *n*
Random initialization of each chromosome in *n*
Set counter = 0
Compute the fitness function of each chromosome
While (counter < *I*)
Select chromosome pair using roulette
Calculate the fitness of the current offspring
Apply crossing over with 70% probability
Apply mutation with 1% probability
Replace old population with new population
Save the current configuration of offspring
Update O_{best}
Increment counter
End while
Return The best solution of configuration O_{best}

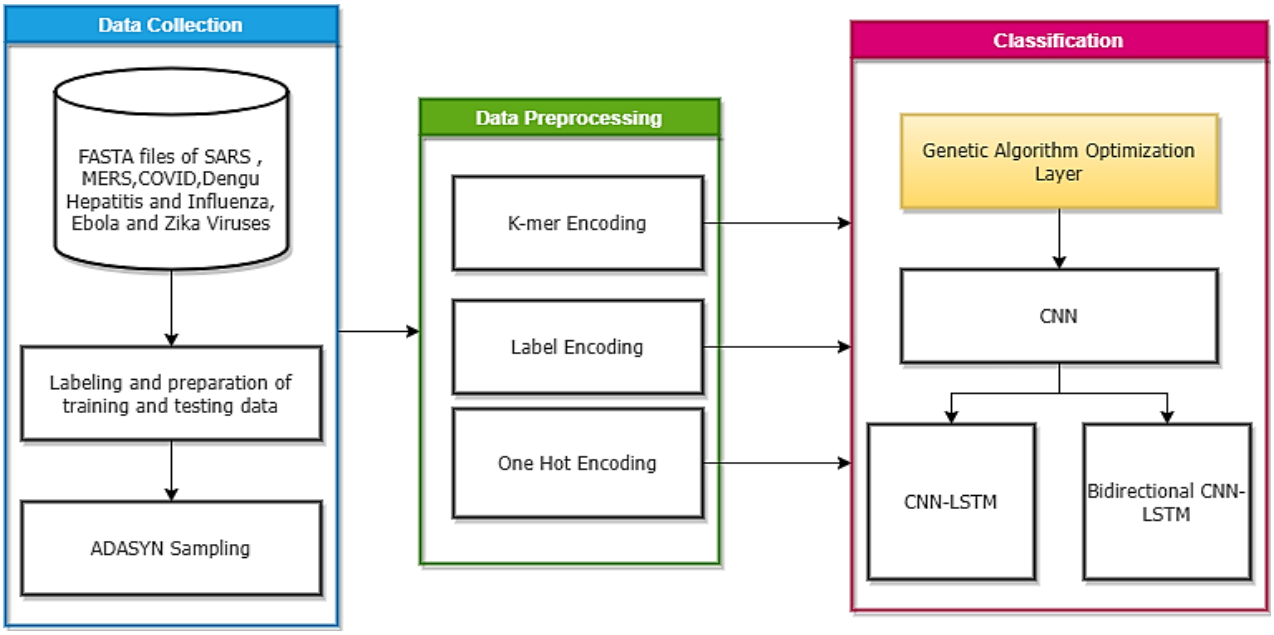


Fig. 3. Summary of the Proposed Workflow.

In the proposed GA algorithm, the chromosomes of the original GA reflect the CNN weights in GA. The population, which is made up of several chromosomes, is seeded at random. The number of weight vectors is represented by the number of chromosomes. The fitness function is the training set's accuracy. As a result, while using CNN, the optimization challenge entails maximizing the accuracy of the training set.

The first phase in the algorithm is the generation of the starting population. This is the first stage of the process. In the CNN model, the values of the hyperparameters are picked at random from the defined search spaces with the help of the python random module, which follows the uniform distribution. The fitness evaluation is the next phase. The validation accuracy and the average of the model's five highest training accuracy were both considered in the trials as fitness functions. The highest accuracy represents the highest fitness. The selection method employed is the roulette wheel. Then, the crossover and mutation stages proceeds. After the crossover occurs, the entire new generation gets mutated. Crossover is accomplished by picking hyperparameters between each parent at random in an equiprobable manner. Additionally, the parents are chosen equitably among the surviving. After forming a new generation, the procedure is repeated iteratively from the second step until the final condition is satisfied. The final condition in the context is the occurrence of a specified number of generations. The output of the algorithm is the configuration of the weight with the highest fitness.

In order to keep track of the GA configuration on each generation after evaluating the loss function, the complete parameters of the generation are saved in memory with its corresponding accuracy as well as the selected parents: a Boolean flag which represents if mutation occurs or not, the mutated individual if any and finally the crossing over Boolean flag.

2) Convolutional neural networks (CNN): In deep learning, Convolutional Neural Networks (CNN) [9] is a commonly used technique that may produce cutting-edge results for the majority of classification problems [9, 28, 29]. CNN not only works well in image classification, but it may also deliver accurate results when dealing with text data. CNN is mostly used to automatically extract the features from an input dataset, as opposed to machine learning models, which need the user to select the features from an input dataset. Text classification is processed using 1D CNN. CNN can only deal with numerical data. Therefore, the DNA sequence must be transformed into numerical values via one-hot encoding or label encoding. The CNN architecture extracts features from the input dataset through the use of a series of convolutional layers. After each convolutional layer, there is a maximum pooling layer, and the size of the derived features is lowered. This layer turns the words into a vector space model based on the frequency with which a word appears near other words in the text. For feature extraction, two convolutional layers with filters of 128 and 64 are used in the model, as well as a kernel of size (2 x 2) with ReLU as the activation function for the extraction of features. A max-pooling layer of size (2x2) is added to the feature map to minimize the overall size of the feature map. The softmax function [30] is utilized as the classification layer. In neural network models that predict a multinomial probability distribution, the softmax function is chosen as the activation function in the output layer. It produces an output that shows the probability of each class label. It can provide good results for multi-classification of DNA sequences. The CNN weights are already optimized due to the previous GA layer. Thus, the accuracy of the produced CNN layer is optimal for using it for the next models.

3) CNN-LSTM and CNN-bidirectional LSTM layers: Long Short-Term Memory (LSTM) [11] is an RNN that can learn the long-term dependencies in a sequence. It is used in the prediction and classification of sequences [10,11,26]. It consists of a succession of cells, each of which has three gates: input, output, and forget. In this situation, the LSTM will only retain certain information and discard others. The LSTM output gate uses the sigmoid activation function and the tanh activation function to analyze the cell state to determine what value can be produced. After the convolutional layers, a 100-memory-unit LSTM layer is added to the model to help predict classification labels. The CNN output features are sent into the LSTM layer for classification. Hybrid models using CNN and LSTM are commonly used in NLP tasks to increase classification accuracy [9,10,11,29,31]. Text classification has been improved by using this hybrid model. With dropout layers and regularization approaches, the overfitting problem is minimized in the LSTM modeling process. DNA sequence classification is performed using a bidirectional LSTM/CNN hybrid model. The model employs a CNN for feature extraction and a bidirectional LSTM for classification. Then, CNN is sent into the Bidirectional LSTM as an input. DNA sequence classification makes use of a bidirectional LSTM/CNN hybrid model. For classification, the model relies on a bidirectional LSTM and CNN. The bidirectional LSTM has two RNNs, one for the forward sequence and one for the backward sequence [32].

IV. Experimental Results

The experiments were conducted on a machine using an NVIDIA 1660Ti GPU processor with a RAM size of 16GB. The CPU of the machine was Intel Core i5-8300H @2.30GHz with 4 Cores and 8 logical processors. The models were trained and tested using Tensorflow [33] in python. The dataset was divided into 60% training, 20% validation, and 20% testing using 10-fold cross-validation.

Before the classification phase, the GA was experimented on with different parameters. Several number of generations to end the GA optimization were used. The best results showed that using 12 generations as the specified number of generations yielded the best results. Several mutation probabilities were also used but the one that yielded the best results was a 1% rate of mutation. The rate of crossing-over used was 70%. The categorical cross-entropy function was used in the case of one hot encoding while binary cross-entropy was used with other embeddings as a loss function in the training phase. The error between the actual output and the goal label, on which the weights are trained and updated, is calculated using the loss function of the GA algorithm. A variety of hyperparameters, such as filter size, layer count, and embedding dimension, were used to evaluate the CNN, CNN-LSTM, and CNN-bidirectional LSTM models but the same architecture is used and the same hyperparameters as [18] in testing and evaluation to correctly compare the results. The embedding layer has 8 dimensions, which is the initial layer. If a word appears often next to other words, this layer transforms it into the vector space. This layer, which employs random weights, is responsible for figuring out how each word in the training dataset should be embedded. For feature extraction, a 2x2 kernel with ReLU as an activation function and two convolutional layers with 128 and 64-bit filters are added to the model. Adding a max-pooling layer of size reduces the feature map dimensions (2x2). Using the flatten layer, the feature maps are finally turned into single-column vectors. A thick layer with neurons 128 and 64 receives the output. The number of filters in each layer are 128, 64, and 32, respectively. The embedding dimension of 32 and a k-mer length of 6 are included in the filter's dimensions. The models were trained with 10 epochs each for each of the encoding methods. The resultant training accuracy for each model is shown in Table II.

The same LSTM and LSTM/CNN hybrid models are used in [18] to correctly compare the results and improve upon the currently existing model after adding the GA layer and using ADASYN for oversampling as well as increasing the dataset variability. Increasing the number of class labels in the dataset and the number of input sequences also contributed to the overall better performance of the models. The accuracy increased as compared to [18] by the introduction of the two new class labels for the Zika and the Ebola virus as well as the additional data collected for the rest of the class labels. Label encoding achieved the best accuracy in the CNN classification layer in both training and testing thus it would achieve the best results in the remaining layers. This is because the CNN layer is used as an input to both the CNN-LSTM layer and the CNN Bidirectional LSTM layer. The models were trained and tested using GA optimization and without using GA optimization. Results show that GA optimization yielded noticeably better results in all label, one-hot and k-mer encodings than the results without GA optimization. Testing results and the results of the experiments using GA optimization and the same experiment without using the optimization layer are shown in Table III.

TABLE II. Training Accuracy of the Proposed Method

ENCODING METHOD	**CLASSIFICATION METHOD**		
	CNN	**CNN-LSTM**	**CNN Bidirectional LSTM**
Label Encoding	**95.12%**	**94.36%**	**93.82%**
One-Hot Encoding	94.57%	93.89%	93.22%
K-mer Encoding	94.51%	94.21%	93.55%

TABLE III. COMPARISON OF CLASSIFICATION MODELS WITH AND WITHOUT GA OPTIMIZATION LAYER USING DIFFERENT ENCODING METHODS

ENCODING METHOD	CLASSIFICATION METHOD					
	CNN		CNN-LSTM		CNN Bidirectional LSTM	
	Using GA Optimization	Without Using GA Optimization	Using GA Optimization	Without Using GA Optimization	Using GA Optimization	Without Using GA Optimization
Label Encoding	**93.51%**	92.92%	**93.27%**	92.78%	**93.20%**	92.14%
One-Hot Encoding	**93.77%**	93.16%	**93.54%**	93.02%	**93.44%**	93.13%
K-mer Encoding	**93.51%**	92.92%	**93.27%**	92.78%	**93.20%**	92.14%

With the addition of the GA optimization layer, label encoding, one hot encoding and k-mer encoding achieved an accuracy of 94.88%, 94.33% and 94.05%, respectively using the CNN model. Using CNN-LSTM, label encoding, one hot encoding and k-mer encoding achieved an accuracy of 94.42%, 93.51% and 93.9%, respectively. Finally utilizing the CNN-LSTM Bidirectional model, the accuracies were 93.74% for label encoding, 93.01% for one-hot encoding and 93.37% for k-mer encoding. On the other hand, without using the GA optimization layer the accuracy for each model was considerably less. CNN achieved an accuracy of 93.22%, 93.50% and 93.54% for label encoding, one hot encoding and k-mer encoding, respectively. Using CNN-LSTM model, label encoding achieved an accuracy of 93.5%, one-hot encoding achieved an accuracy of 91.59% and k-mer encoding showed an accuracy of 92.16%. Finally, CNN Bidirectional model achieved an accuracy of 91.35%, 92.16% and 92.46% for label encoding, one hot encoding and k-mer encoding, respectively. Among all the three encoding techniques, label encoding is shown to achieve the best results overall with the introduction of the GA layer and without using it.

In order to compare the results with [18], the two additional class labels Zika and EBOLA viruses were removed from the dataset and then the dataset was experimented on. Thus, the experiment was carried on using ADASYN for oversampling and the addition of the GA optimization layer in comparison with [18] who used SMOTE and the hybrid model without the addition of the GA layer. The resultant accuracies are shown in Table IV. Furthermore, only label encoding and k-mer encoding is demonstrated for comparison as in [18].

By comparing the results of k-mer encoding using GA and introducing two new class labels to the dataset and ADASYN oversampling method, the proposed method is proved to give better accuracy results than the previous model used by [18]. The best results from [18] were achieved using k-mer encoding. In the proposed method in this study the resulting accuracy using k-mer encoding were 94.05% using CNN, 93.9% using CNN-LSTM and 93.37% using CNN Bidirectional LSTM. Whereas it previously resulted in 93.16% using CNN, 93.02% using CNN-LSTM and 93.13% using CNN-Bidirectional LSTM without GA optimization and using SMOTE oversampling with less dataset sequences and less class labels. Thus, the proposed method achieved best accuracy using k-mer encoding in comparison to [18]. It also achieved the best overall classification accuracy of 94.88% using label encoding. The training and validation loss curves for the three encoding methods are shown in Fig. 4.

The accuracy curve shows that label encoding achieved the best training and testing results overall among all the three used encoding methods. One hot encoding showed similar results for both training and testing in CNN Bidirectional LSTM but better training accuracy using CNN and LSTM. Utilizing ADASYN resulted in better results in the overall training accuracy due to the optimized oversampling of the dataset in the minority class labels such as Zika and Dengue. As a limitation, improving the accuracy by introducing the optimization layer leads to an increase in computational time. Moreover, the generated synthetic dataset in the oversampling method might have some fuzzy class boundaries.

TABLE IV. COMPARISON OF CLASSIFICATION MODELS WITH GUNASEKARAN, ET AL. [18] USING GA OPTIMIZATION AND WITHOUT THE ADDITION OF THE 2 NEW CLASS LABELS

ENCODING METHOD	CLASSIFICATION METHOD					
	CNN		CNN-LSTM		CNN Bidirectional LSTM	
	Proposed Model	Gunasekaran, et al. [18]	Proposed Model	Gunasekaran, et al. [18]	Proposed Model	Gunasekaran, et al. [18]
Label Encoding	**93.51%**	92.92%	**93.27%**	92.78%	**93.20%**	92.14%
K-mer Encoding	**93.77%**	93.16%	**93.54%**	93.02%	**93.44%**	93.13%

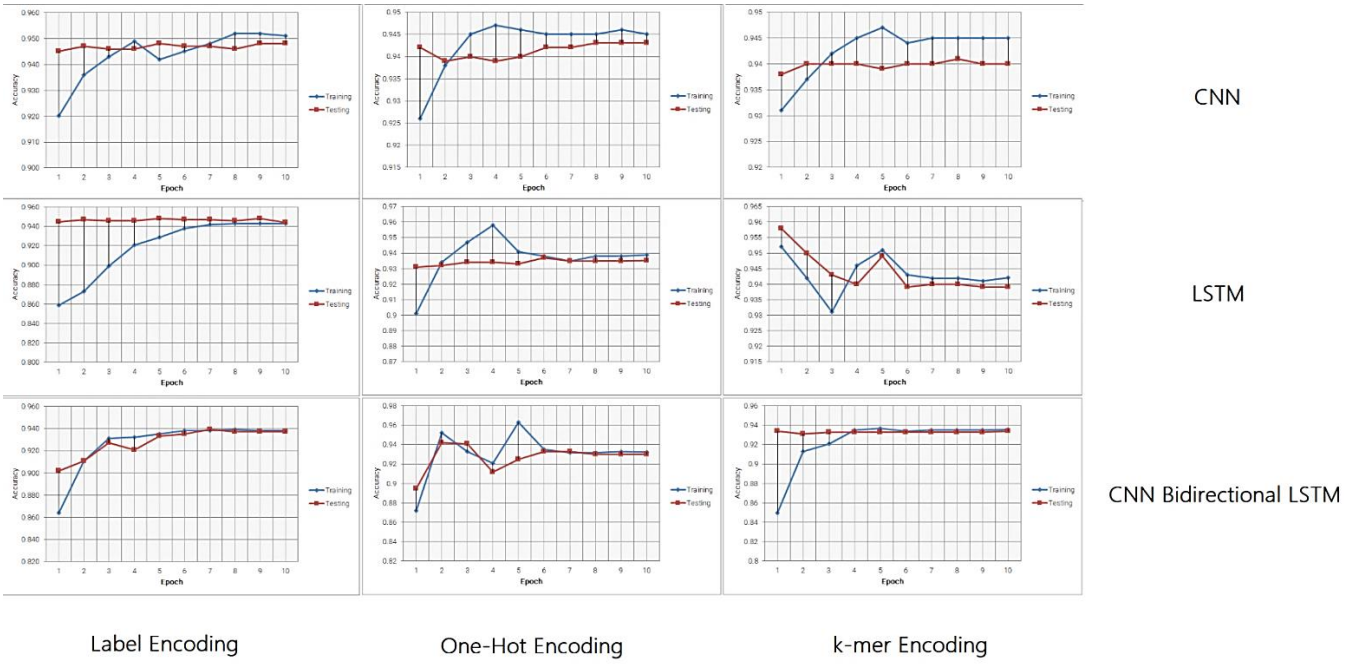


Fig. 4. The Resultant Training and Validation Loss Curves using GA Optimization and without using GA Optimization with all 3 Different Encoding Methods.

V. Conclusion

The classification of viral DNA poses a major challenge in recent years. The accurate classification of the DNA of pandemic viruses will greatly help in the production of vaccines and the identification of new pathogens. This study proposes an optimized method for the accurate classification of viral DNA utilizing genetic algorithm for optimization classification using a hybrid deep learning model. The proposed method uses a genetic algorithm to optimize the weights of the CNN model which enhances the overall classification accuracy. The study also utilizes ADASYN as an optimized dataset oversampling technique for the minority class labels. Three encoding techniques were experimented with which are label encoding, k-mer encoding, and one-hot encoding which was not used in previously proposed models. The experiments showed that the proposed optimization layer GA and ADASYN with the deep learning model outperformed previously proposed models on the same dataset in terms of classification accuracy. The models were then trained and tested with GA optimization and without GA optimization. The GA optimization drastically affected the accuracy of the models. As a result, label encoding was shown to achieve the best accuracy of 94.88% using CNN. Besides, k-mer encoding achieved an accuracy of 94.05% whereas the best results achieved by a previously proposed model were 93.16%. As a result, it is shown that the introduction of an optimization layer improved the overall classification accuracy. The introduction of more evolutionary or optimization algorithms in future research could improve the accuracy further. Furthermore, the use of an optimized oversampling technique yielded better overall accuracy. Therefore, by using ADASYN which is an optimized version of SMOTE yielded better results.

For future work, it is planned to introduce more viral DNA sequences in the training dataset and use other selection criteria for the GA selection algorithm which could further improve the accuracy of the classification. In addition, more optimization methods could be utilized.

References

[1] Trovato, M., Sartorius, R., D'Apice, L., Manco, R. and De Berardinis, P., 2020. Viral emerging diseases: challenges in developing vaccination strategies. Frontiers in Immunology, 11, p.2130.

[2] Gelderblom, H.R., 1996. Structure and classification of viruses. Medical Microbiology. 4th edition.

[3] Travers, A. and Muskhelishvili, G., 2015. DNA structure and function. The FEBS journal, 282(12), pp.2279-2295.

[4] Randhawa, G.S., Soltysiak, M.P., El Roz, H., de Souza, C.P., Hill, K.A. and Kari, L., 2020. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. Plos one, 15(4), p.e0232391.

[5] Liu, R., Qiao, M., Zheng, J. and Zhou, W., 2021. Analysis SARS-CoV-2 Genomes of G20 Areas on Phylogeny Tree, t-SNE based on Machine Learning.

[6] Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y. and Zhang, L., 2020. Review on the application of machine learning algorithms in the sequence data mining of DNA. Frontiers in Bioengineering and Biotechnology, 8, p.1032.

[7] Abd–Alhalem, S.M., El-Rabaie, E.S.M., Soliman, N., Abdulrahman, S.E.S., Ismail, N.A., El-samie, A. and Fathi, E., 2021. DNA Sequences Classification with Deep Learning: A Survey. Menoufia Journal of Electronic Engineering Research, 30(1), pp.41-51.

[8] Deorowicz, S., 2020. FQSqueezer: k-mer-based compression of sequencing data. Scientific reports, 10(1), pp.1-9.

[9] Nguyen, N.G., Tran, V.A., Phan, D., Lumbanraja, F.R., Faisal, M.R., Abapihi, B., Kubo, M. and Satou, K., 2016. DNA sequence classification by convolutional neural network. Journal Biomedical Science and Engineering, 9(5), pp.280-286.

[10] Sharma, A., Lysenko, A., Boroevich, K.A., Vans, E. and Tsunoda, T., 2021. DeepFeature: feature selection in nonimage data using convolutional neural network. Briefings in bioinformatics, 22(6), p.bbab297.

[11] Nowak, J., Taspinar, A. and Scherer, R., 2017, June. LSTM recurrent neural networks for short text and sentiment classification. In

International Conference on Artificial Intelligence and Soft Computing (pp. 553-562). Springer, Cham.

[12] Katoch, S., Chauhan, S.S. and Kumar, V., 2021. A review on genetic algorithm: past, present, and future. Multimedia Tools and Applications, 80(5), pp.8091-8126.

[13] Abd-El-Wahed, W.F., Mousa, A.A. and El-Shorbagy, M.A., 2011. Integrating particle swarm optimization with genetic algorithms for solving nonlinear optimization problems. Journal of Computational and Applied Mathematics, 235(5), pp.1446-1453.

[14] Hamdia, K.M., Zhuang, X. and Rabczuk, T., 2021. An efficient optimization approach for designing machine learning models based on genetic algorithm. Neural Computing and Applications, 33(6), pp.1923-1933.

[15] Idrissi, M.A.J., Ramchoun, H., Ghanou, Y. and Ettaouil, M., 2016, May. Genetic algorithm for neural network architecture optimization. In 2016 3rd International conference on logistics operations management (GOL) (pp. 1-4). IEEE.

[16] Asim, M.N., Malik, M.I., Dengel, A. and Ahmed, S., 2020, July. K-mer Neural Embedding Performance Analysis Using Amino Acid Codons. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[17] Humayun, F., Khan, F., Fawad, N., Shamas, S., Fazal, S., Khan, A., Ali, A., Farhan, A. and Wei, D.Q., 2021. Computational method for classification of avian influenza A virus using DNA sequence information and physicochemical properties. Frontiers in Genetics, 12, p.599321.

[18] Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Deepa Kanmani, S., Venkatesan, C. and Suresh Gnana Dhas, C., 2021. Analysis of DNA sequence classification using CNN and hybrid models. Computational and Mathematical Methods in Medicine, 2021.

[19] Bzhalava, Z., Tampuu, A., Bała, P., Vicente, R. and Dillner, J., 2018. Machine Learning for detection of viral sequences in human metagenomic datasets. BMC bioinformatics, 19(1), pp.1-11.

[20] Syahrani, I.M., 2019. Comparation analysis of ensemble technique with boosting (Xgboost) and bagging (Randomforest) for classify splice junction DNA sequence category. Jurnal Penelitian Pos dan Informatika, 9(1), pp.27-36.

[21] He, H., Bai, Y., Garcia, E.A. and Li, S., 2008, June. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). IEEE.

[22] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357.

[23] Marques, Y.B., de Paiva Oliveira, A., Ribeiro Vasconcelos, A.T. and Cerqueira, F.R., 2016. Mirnacle: machine learning with SMOTE and random forest for improving selectivity in pre-miRNA ab initio prediction. BMC bioinformatics, 17(18), pp.53-63.

[24] Miao, Y., Liu, F., Hou, T. and Liu, Y., 2022. Virtifier: a deep learning-based identifier for viral sequences from metagenomes. Bioinformatics, 38(5), pp.1216-1222.

[25] Loussaief, S. and Abdelkrim, A., 2018. Convolutional neural network hyper-parameters optimization based on genetic algorithms. International Journal of Advanced Computer Science and Applications, 9(10).

[26] Chen, M., Yu, L., Zhi, C., Sun, R., Zhu, S., Gao, Z., Ke, Z., Zhu, M. and Zhang, Y., 2022. Improved faster R-CNN for fabric defect detection based on Gabor filter with Genetic Algorithm optimization. Computers in Industry, 134, p.103551.

[27] Bhandari, A., Tripathy, B.K., Jawad, K., Bhatia, S., Rahmani, M.K.I. and Mashat, A., 2022. Cancer Detection and Prediction Using Genetic Algorithms. Computational Intelligence and Neuroscience, 2022.

[28] Aoki, G. and Sakakibara, Y., 2018. Convolutional neural networks for classification of alignments of non-coding RNA sequences. Bioinformatics, 34(13), pp.i237-i244.

[29] Min, S., Lee, B. and Yoon, S., 2017. Deep learning in bioinformatics. Briefings in bioinformatics, 18(5), pp.851-869.

[30] Bridle, J., 1989. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. Advances in neural information processing systems, 2.

[31] Lugo, L. and Hernández, E.B., 2021. A Recurrent Neural Network approach for whole genome bacteria identification. Applied Artificial Intelligence, 35(9), pp.642-656.

[32] Mughees, N., Mohsin, S.A., Mughees, A. and Mughees, A., 2021. Deep sequence to sequence Bi-LSTM neural networks for day-ahead peak load forecasting. Expert Systems with Applications, 175, p.114844.

[33] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. and Kudlur, M., 2016. {TensorFlow}: a system for {Large-Scale} machine learning. In 12th USENIX symposium on operating systems design and implementation (OSDI 16) (pp. 265-283).

J-Selaras: New Algorithm for Automated Data Integration Tools

Mustafa Man[1]*
Faculty of Ocean Engineering Technology and Informatics
Universiti Malaysia Terengganu (UMT)
21030 Kuala Nerus, Terengganu, Malaysia

Mohd. Kamir Yusof[3]
Faculty of Informatics and Computing
Universiti Sultan Zainal Abidin (UniSZA)
Besut Campus, 22200 Besut, Terengganu, Malaysia

Wan Aezwani Wan Abu Bakar[2]*
Pusat Asasi Sains dan Perubatan,
Universiti Sultan Zainal Abidin (UniSZA)
Gong Badak Campus, 21030 Kuala Nerus, Terengganu, Malaysia

Norisah Abdul Ghani[4], Mohd Adza Arshad[5]
JKR Centre of Excellence for Engineering and Technology (CREaTE)
Public Work of Department (JKR) Malaysia, Jalan Kemus, Simpang Ampat, 78000 Alor Gajah, Melaka, Malaysia

Raja Normawati Raja Ayob[6], Kamarul Azhar Mahmood[7], Faizul Azwan Ariffin[8], Mohamad Dizi Che Kadir[9], Lily Mariya Rosli[10], Nurhafiza Binti Abu Yaziz[11]
Cawangan Kontrak dan Ukur Bahan (CKUB)
Public Work of Department (JKR) Malaysia, Tingkat 16, Menara Tun Ismail Mohamed Ali, No. 25, Jalan Raja Laut 50350 Kuala Lumpur, Malaysia

***Abstract*—Data integration is a popular technique or method today for data converting and sharing within new application with different database format and location. The interaction of data from one application system to another application system requires an intermediary software or middleware that allows the data to be transferred or read systematically and easily. The development of dynamic algorithms allows data in various formats, whether structured or unstructured, to be transferred to various types of databases smoothly. A case study was conducted for the Bill of Quantity (BQ) data in the known Excel format generated through CostX software in a single sheet Excel file. It was transferred to a single workbook with multiple sheets with formulation generated automatically. Thus, an algorithm was developed and tested through the development of the J-Selaras System. This algorithm can remove the noisy data or data symbols that are not related in the excel single sheet (CostX) file and automatically transfer to multiple excel sheets with macros formulation quickly. The implementation results indicate a significant contribution where it reduces in execution time of BQ processes and manpower resources used.**

***Keywords*—Automated data integration algorithm; bill of quantity (BQ); CostX; single and multiple sheets; J-Selaras tool**

I. Introduction

This research case study focuses on Malaysian government agency, named as Jabatan Kerja Raya (JKR) which is responsible for the BQ for construction in a tendering process. The collaboration is done to solve the manual process in converting The CostX single sheet BQ into MS Excel multiple sheets with macro function.

The Malaysian Public Works Department or simplified as Jabatan Kerja Raya (JKR) Malaysia is the Malaysia's federal government department under the roof of Ministry of Works Malaysia (MOW), that is accountable for construction and maintenance activities on public facility and infrastructure [1]. The JKR Strategic Plan 2021-2025 was issued, and its outlines are five strategic themes. The initiatives in the strategic themes are to realize the government effort in achieving the Shared Prosperity Vision 2030. In the alignment to the first theme in the JKR Strategic Framework, the demands for a project delivery in terms of the completion within the stipulated time, within the budget and the stated quality becomes more prominent in fulfilling customer satisfaction.

There is the need to expedite the preparation of the contract document after signing of the Letter of Acceptance (LoA) by the contractor before the project is being awarded. Both the Government and the contractor must fulfil their obligations according to the contract's provisions. Typically, the contractor must provide the contract conditions, the tender drawings, the Bills of Quantities (BQ), and the other related documents throughout the process of tendering. These materials will be bound into the Contract Document after the LoA is issued. The contract document must be prepared within four (4) months upon signing of the LoA by the Contractor [2]. A Bill of Quantities (BQ) [3] is a document created by a quantity surveyor or cost consultant that contains information on certain sections. The sections may include Form of Tender, Information such as the scope of work, Requirements on certain material's quantity, Pricing schedule, Provisional sums, and Day works and the labor costs in a construction project. The BQ is an essential part of the tender. Without a suitable BQ, a tender is incomplete. The contractors can use this information and data to quote rates for their specific work.

*Corresponding Author

The purpose of BQ is to make the tendering process more uniform. Other objectives are to establish a transparent and exact method for valuing the project, to provide a thorough description of the work and its rates, as well as the overall cost. A solid BQ helps quantity surveyors in ensuring that individual contractors have filed valid tenders that comply with the specifications.

The CostX [4] is the offline database system using PostgreSQL. CostX is a Construction Management Software that is designed to serve a construction activity especially for Enterprises or Small Medium Enterprise company. The limitation of CostX is subject to the offline system and it provides the details of each itemize sections for construction in only single sheet. This research aims at solving the manual process of transferring BQ from CostX system into Excel format by providing the automated converting tool for BQ to ease the price generation during tendering activities. The delay in preparing the contract document is due to the lengthy time taken in the process of the Rationalizing of Rates for Bills of Quantities (RoBQ) which are the part of the contract document. The RoBQ needs to be carefully and thoroughly examined, the reasonableness of the tenderer price rates, and the reasonably adjusted price rates will become the BQ contract rates. The BQ contract rates will determine the value of the variation work if there are any variation during the construction period. If any, the value of the variation work will increase the cost of the project delivery and will affect the initial budget approved. Currently, in JKR, the RoBQ process is not yet adopting an information and Communication Technology (ICT), and therefore the process of the RoBQ will be time consuming.

Data integration [5] is the data combination from different sources to help data managers and executives analyze it and make smarter business decisions. In this project, we deploy integration of formatting in CostX system via PostgreSQL database [6] and MS Excel format into a structured MySQL database in generating standardized BQ pricing with a verification and validation from an accountable party for the user in construction company.

II. Proposed Methodology

The process of tendering process comprises of six steps as illustrated in Fig. 1. Step 1 consists of the filling up the BQ using CostX with PostgreSQL database [6] system and converted to Excel macro format that is stored in Cloud database. In step 2, the softcopy of BQ estimation in Excel format is given to the successful tenderers and shortlisted tenderers that is obtained from the tender appraiser officer. Then in step 3, the BQ in Excel format from the estimation department that is awarded to the successful tenderers is uploaded to the J-Selaras system. In step 4, J-Selaras system will create a pricing comparison table. Then in step 5, each price rates for each BQ will be analyzed using the cut-off concepts such as mean, max, average, and standard deviation figures. The adjustment price prior to the cut-off formula will be reviewed by the appointed reviewer and the verified officer in charge in step 6. The J-Selaras system will check for arithmetical error and the adjusted price is matched with the actual price in the contract tender as stated in the Letter of Acceptance (SST).

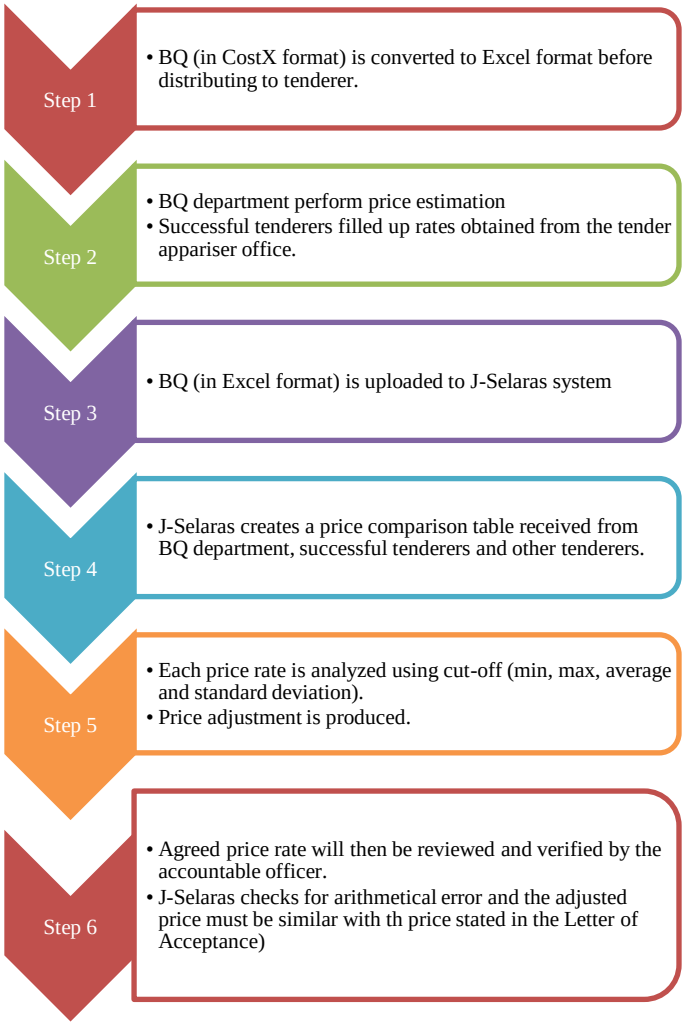


Fig. 1. Steps Involved in Data Transfer Process from CostX to Excel Format and the BQ Tender Price Adjustment Rate Determination Process.

III. J-Selaras Development Model

The J-Selaras converting tool is developed based on filtering algorithm as depicted in Fig. 2.

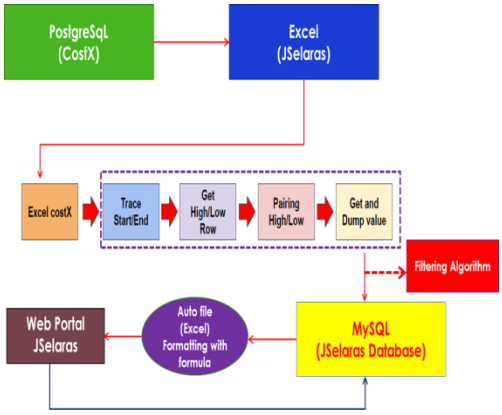


Fig. 2. J-Selaras Model.

The detail specification of BQ is first filled up into CostX system that is built with PostgreSQL database. Then, from CostX system, the BQ specification is downloaded in excel format and converted into J-Selaras system through filtering algorithm by first, tracing the Start and End (Start/End) keywords. Secondly, the cells that cover from Start to the End will be highlighted to get the High row to Low row. Thirdly, those cells involved in High and Low rows will be paired before getting all values involved in all the cells. Refer to next section for details on Filtering algorithm.

A. Filtering Algorithm

Fig. 3 shows the steps in filtering algorithm where the conversion is done from CostX file format into Excel format.

Pre-requisite: $\forall s \{s: each\ sheets\ in\ Excel\}$
Start
1. *Trace the Start and End Keyword*
2. *Focus the value from the highest row to the lowest row*
3. *Pair the value from the highest row to the lowest row*
4. *Get the value and dump into the MySQL database of J-Selaras system*

End

Fig. 3. Filtering Algorithm.

B. BQ Pricing Standardization

Prior to being processed through filtering algorithm, the Bill of Quantity (BQ) is now converted and saved into MySQL database in J-Selaras system in Cloud environment. The list of BQ price is not yet standardized. The generated price in BQ needs to follow the standardization price that is set by the construction company. All prices of the materials set in BQ is standardized according to certain formula such as Z-Score [7]. The formula involved in Z-Score are Standard Deviation (SD), Min, Max, and cut-off indicates the pseudocode for Z-Score formula used in BQ pricing. The pseudocode for a BQ applying the Z-Score formulation is depicted in Fig. 4.

Given the BQ price of items, then
Start
1. *Compute the rMin and rMax*
2. *Analyse the price with cut-off*
3. *J-Selaras propose the BQ price*
   - *3.1 If the proposed BQ price > 115% and < +10%*
     - *3.1.1 J-Selaras accept the proposed price*
   - *3.2 Else*
     - *3.2.1 J-Selaras accept the cut-off price*

End
End

Fig. 4. Pseudocode in Z-Score Formulation.

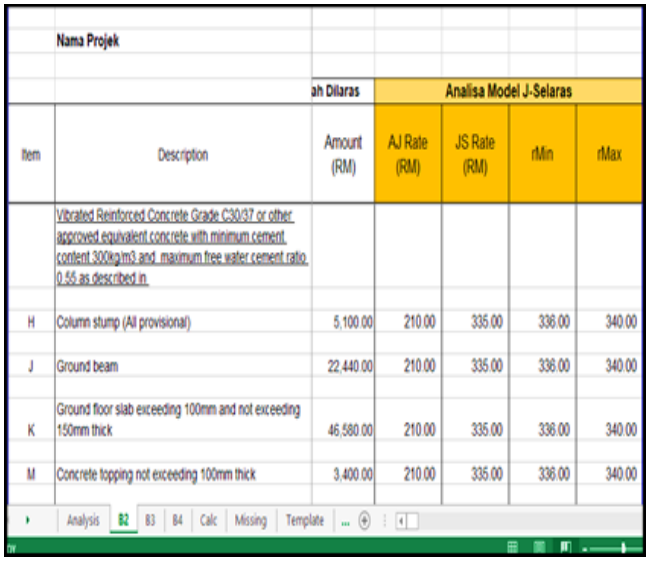
Z-score [8-9] is based on the calculation of mean and standard deviation of an attribute. It is a measurement of difference between individual value and the mean population, and then divided by standard deviation of population. The computed, Z-score (Z) provides each feature with a zero mean and a unit variance. The foundation of Z-score is where the mathematical Gaussian curve or 'Bell Shaped' curve is applied to the data under study [10-11]. The Z-score, Z as in [12] is expressed as follow:

$$\text{Z- Score} = \frac{x_i - \bar{x}}{s} \quad (1)$$

Where x_i is an individual value, $\bar{x}$ is mean of samples and s is standard deviation of samples. The Z-score technique was proposed in the study to analyze the comparison of each rate in the BQ between all the short-listed tenderer, the successful tenderer and the department's cost estimation [13]. Meanwhile, the proposed rates in the BQ to be agreed by the successful tenderer will be automatically generated by the system based on the cut-off formula i.e., the rate of an item description is derived from the average rates of the total number of all the short-listed tenderer including the rate of the successful tenderer and the rate of the department's estimate [14]. The cut-off formula and the Z-score technique are like the tender evaluation system format. The cut-off principle means the lowest acceptable rate to be certified. The tenderer will not be able to complete the project if the rate is too low i.e., the rate is lower than and below the cut-off.

In accordance with statistical methods, the setting of the cut-off rate is based on 'Mean' and 'Standard Deviation' of the shortlisted tenderer rate including the Department's Estimates (AJ) rates, after rate 'freak' is removed. The range of the cut-off rate is in between not exceeded above +10% (<+10%) and is not exceeded below -15% (<-15%). After all the tenderer price BQ have been uploaded in the J-Selaras, the system will automatically do the analysis based on the setting of the cut-off and the z-score formula in the J-Selaras. Simultaneously, the J-Selaras suggests the contractor rate if the rate is within the cut-off range or suggests the cut-off rate if the contractor rate is out of the range.

In finalizing BQ for the contract document, the appropriate choice of the reasonable rate can be ultimately determined whether to accept the rate of a successful tenderer or the rate generated by the system or the rate of past projects or the current market review rate [15]. The reporting of the analysis is in the tabulation as shown in the Fig. 5. The added value, the J-Selaras system can be as the data cost collection center for price rate and as a reference especially in preparing department's estimate for new project.



Nama Projek

		ah Dilaras	Analisa Model J-Selaras			
Item	Description	Amount (RM)	AJ Rate (RM)	JS Rate (RM)	rMin	rMax
	Vibrated Reinforced Concrete Grade C30/37 or other approved equivalent concrete with minimum cement content 300kg/m3 and maximum free water cement ratio 0.55 as described in					
H	Column stump (All provisional)	5,100.00	210.00	335.00	336.00	340.00
J	Ground beam	22,440.00	210.00	335.00	336.00	340.00
K	Ground floor slab exceeding 100mm and not exceeding 150mm thick	46,580.00	210.00	335.00	336.00	340.00
M	Concrete topping not exceeding 100mm thick	3,400.00	210.00	335.00	336.00	340.00

Analysis | B2 | B3 | B4 | Calc | Missing | Template | ...

Fig. 5. Illustration of the Comparison and Analysis Table.

IV. Experimentation Result

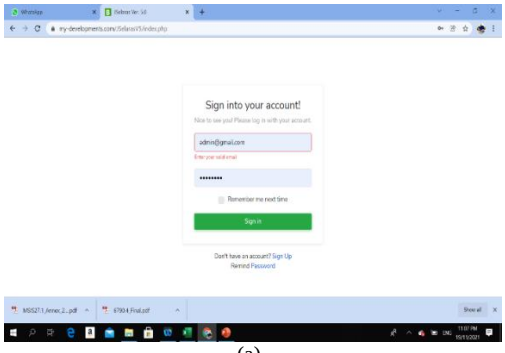
The J-Selaras system is developed under machine specification of VivoBook_ASUS Laptop X415EA_A416EA, Windows 11, i5-1135G7 and 12GB RAM as in Table I.

TABLE I. Machine Specification

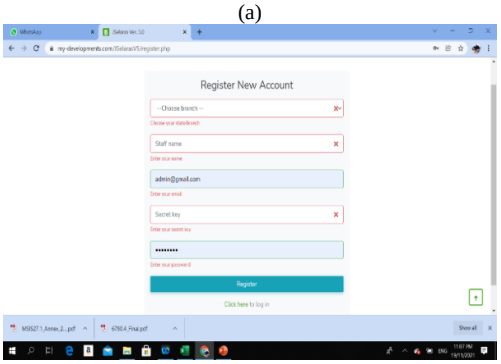
System Model:	VivoBook_ASUS Laptop X415EA_A416EA
Operating System:	Windows 11 Home Single Language 64-bit (10.0, Build 22000) (22000.co_release.210604-1628)
Processor:	11th Gen Intel(R) Core (TM) @ 2.40GHz (8 CPUs), ~2.4GHz
Memory:	12288MB RAM
Available OS Memory:	11982MB RAM

A. Graphical User Interface

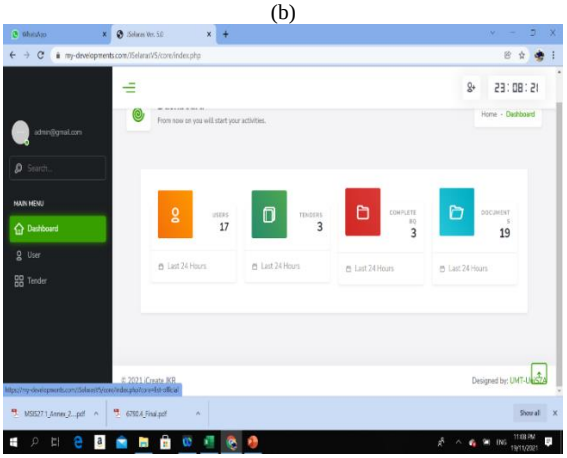
J-Selaras consists of seven modules i.e., Sign-up/Login-Logout Module, Registration of User Profile module, Upload-Download of CostX file format module, Excel Standardization format module, Analysis of Price-rate module, Report generation module and Maintenance module. The J-Selaras system is accessible at https://my-developments.com/JSelarasV5. Fig. 6(a) depicts the login page of J-Selaras admin while Fig. 6(b) shows the new user registration of J-Selaras system. Once the registration successful, then user is directed to J-Selaras dashboard as shown in Fig. 6(c). Fig. 6(d) indicates the list of J-Selaras users that are currently active while Fig. 6(e) and Fig. 6(f) portray all J-Selaras *admin* users and *Pentadbir J-Selaras* users. Fig. 6(g) to Fig. 6(h) dictate the Darul Quran list of projects and sample of Darul Quran project respectively. Meanwhile, Fig. 6(i) states the BQ that is ready to be downloaded and edited the price and Fig. 6(j) shows the sample of generated BQ price. Finally, Fig. 6(k) illustrates the sample of summarized BQ by J-Selaras.



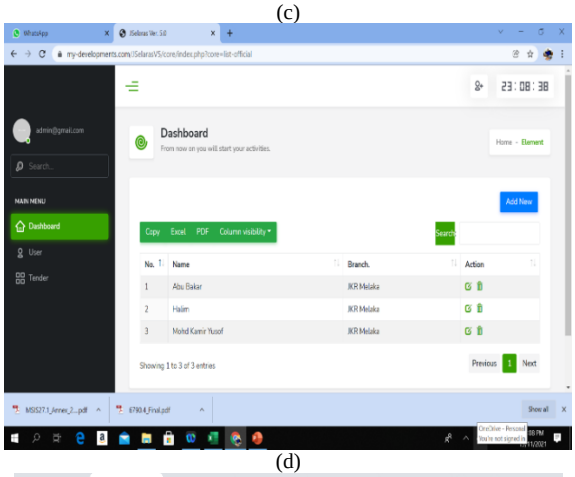
(a)



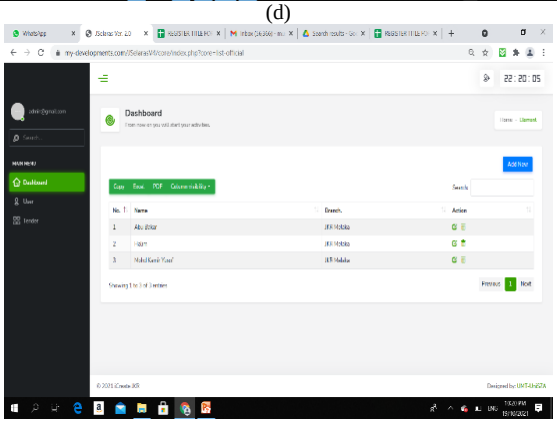
(b)



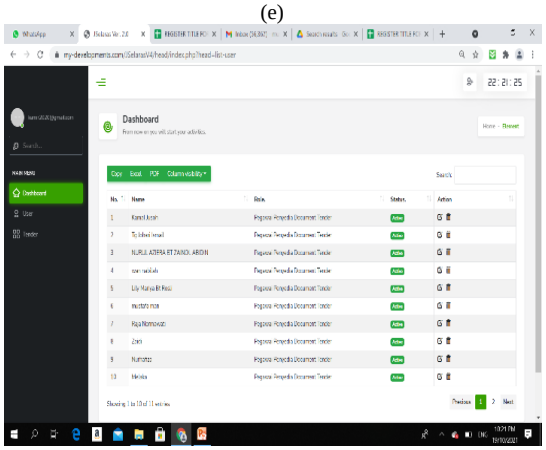
(c)



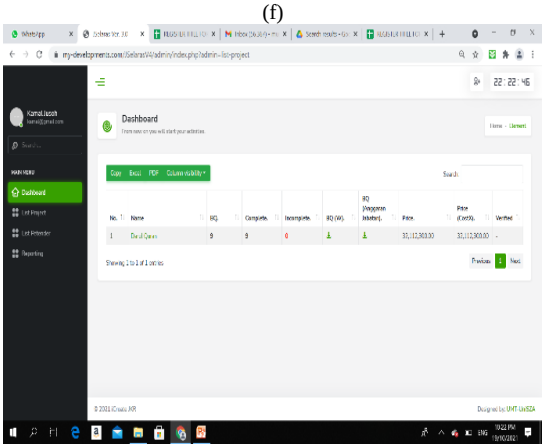
(d)



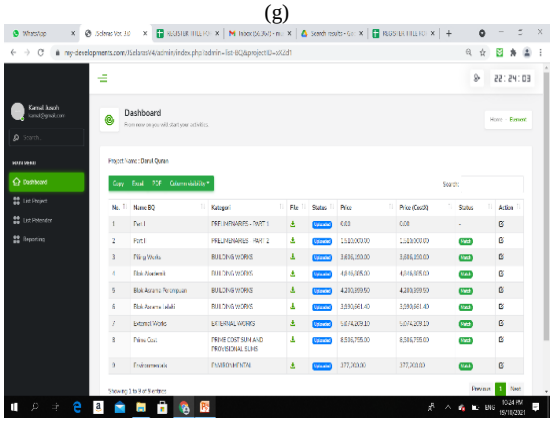
(e)



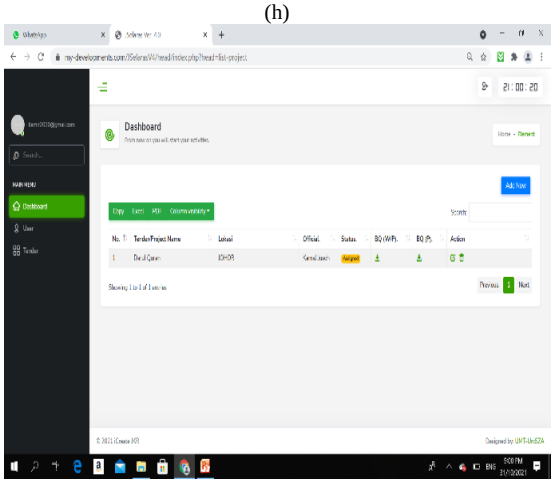
(f)



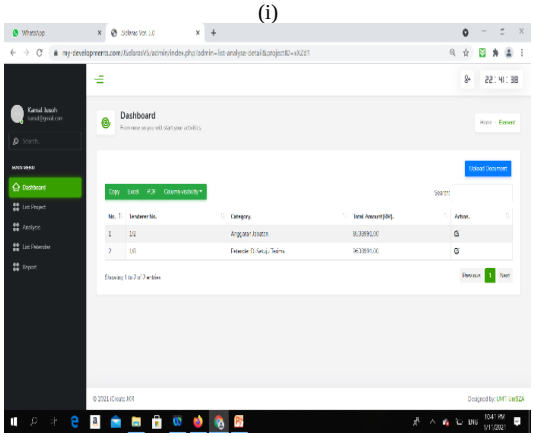
(g)



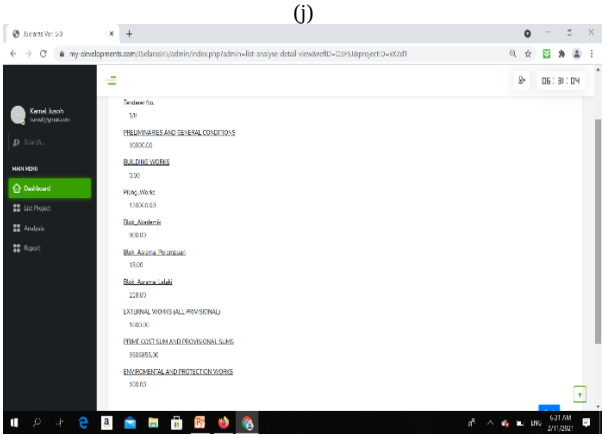
(h)



(i)



(j)



(k)

Fig. 6. a). J-Selaras Admin Login, b). New Account Registration, c). J-Selaras Dashboard, d). List of J-Selaras Users, e). List of User (Admin), f). List of User (Pentadbir J-Selaras), g). List of Projects (Darul Quran), h). Sample of Details BQ of Darul Quran, i). BQ that is Ready to be Downloaded and Edit the Price, j). Sample of BQ Estimated Price Generated, k). Sample of Summarized BQ.

B. *Simulation Result*

Table II plots the result of simulation on BQ generation between the manual processes with the automated J-Selaras system.

TABLE II. THE EVALUATION OF THE EXECUTION TIME IN SECOND (S) BETWEEN THE MANUAL BQ GENERATION WITH J-SELARAS BQ GENERATION

Project	File Size (KB)	Manual Time Process (Ms Excel)(s)	Jselaras Tools time Process (s)
BQ for Project A	297	18000	3
BQ for Project B	301	21600	4
BQ for Project C	294	18000	3
BQ for Project D	424	25200	4
BQ for Project E	600	27000	5

The difference of time execution is recorded as the performance measurement in comparing the manual process of using CostX system with the proposed J-Selaras system for BQ generation of Project A to Project B. The processing time is recorded in second and the results from Table II is plotted in Fig. 7.

The recorded data is based on the five projects (Project A, Project B, Project C, Project D and Project E), the Excel file size in kilobyte (KB) and the execution time processing of the five projects. The manual process enquires more than 18000 seconds as compared to J-Selaras system that generates the BQ from three to five seconds only. With automated conversion tool as J-Selaras, we may have 6K times reduction in generating the BQ.

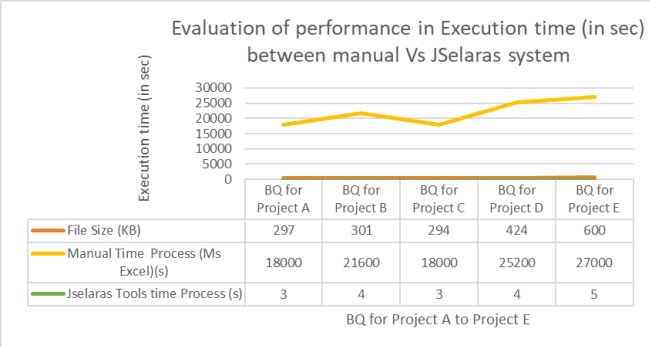


Fig. 7. The Execution Time of BQ Generation between Manual Process using CostX Versus Proposed J-Selaras System.

V. Conclusion

The proposed J-Selaras system has improved the time efficiency and eased the manual conversion done by the manpower. Despite the use of this converter (J-Selaras), the degree of error-rate might be reduced because of the automated conversion being done from the source file of Excel format in CostX system. Moreover, once the data is recorded and saved in centralized database such as MySQL in cloud database platform [16-17], the record keeping mechanism is structured for easy saving and retrieving at the admin own pace. The cloud database storage perhaps is resistant to difficulties in maintenance. The J-Selaras system may act as a construction tool which covers all activities in construction management to serve both top-level management as well as the operation level decision makers.

Acknowledgment

We thank to all entity for the contribution towards this paper writing and publications especially Sr. Norisah Abdul Ghani and Sr. Mohd Adza Arshad from Pusat Kecemerlangan Kejuruteraan dan Teknologi Jabatan Kerja Raya Malaysia (CREaTE), Sr. Raja Normawati Raja Ayob, Sr. Kamarul Azhar Mahmood, Sr. Faizul Azwan Ariffin, Sr. Mohamad Dizi Che Kadir, Sr. Lily Mariya Rosli, Pn. Nurhafiza binti Abu Yaziz from Cawangan Kontrak dan Ukur Bahan (CKUB), JKR Malaysia, Associate. Prof. Ts. Dr. Mustafa Man from Universiti Malaysia Terengganu (UMT) for the idea creation and the conceptual design and implementation, Dr. Mohd. Kamir Yusof and Dr. Wan Aezwani Wan Abu Bakar from Universiti Sultan Zainal Abidin (UniSZA) for the system development and system documentation.

References

[1] L.H. Aun. "Malaysia's Shared Prosperity Vision 2030 Needs a Rethink to Make a Breakthrough." 2019.

[2] Jabatan Kerja Raya Malaysia (JKR). Strategic Plan 2021-2025 from http://www.jk.gov.my. Accessed: 23/08/2022.

[3] P. A. A. Permadi, N. I. W. Oei, and B. Hasiholan. "Comparative Study in Bill of Quantity Estimates on Reinforcement Works of Pile Cap, Single Pier and Double Pier of Flyover Between Conventional Methods and BIM (Building Information Modelling)." *IOP Conference Series: Earth and Environmental Science*. Vol. 1065. No. 1. IOP Publishing, 2022.

[4] iTWO costX Estimating Software from https://www/itwocostx.com/. Accessed: 20/07/2022.

[5] Man, Mustafa, Julaily Aida Jusuh, Mohd Shafry Mohd Rahim, and Mohammad Zaidi Zakaria. "Formal specification for spatial information databases integration framework (SIDIF)." *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 9, no. 1, 2011, pp. 81-88.

[6] What is PostgreSQL? -Definition from Techopedia from https://www.technopedia.com/definition. Accessed: 20/07/2022.

[7] B. Emilian, and C. Darie. "Beginning PHP and MySQL E-Commerce: From Novice to Professional, (Beginners/Beginning Guide)." 2008.

[8] D. Chappel, "Professional Practice for Architects and Project Managers", Wiley, 2020, pp.183-185, ISBN: 978-1-119-54007-6.

[9] S. Sudirman, Z. M. Yusof. "Public sector ict strategic planning: framework of monitoring and evaluating process." Asia-Pacific Journal of Information Technology and Multimedia, 2017, 6(1). pp. 85-99.

[10] JKR Rate On-Line System (RATOL), from ratol.jkr.gov.my. Accessed: 20/07/2022.

[11] R. W. Thatcher, C. J. Biver, & D. M. North. "Z-Score EEG Biofeedback: Technical Foundations." Applied Neurosciences Inc., 2004.

[12] T. C. Brown, T. C. Daniel, & R. M. Forest. "Scaling of ratings: concepts and methods.", 1990.

[13] P. Davis, and D. Baccarini. "The use of bills of quantities in construction projects-an Australian survey." In Proceedings of the COBRA 2004 International Construction Research Conference of the Royal Institution of Chartered Surveyors. RICS Foundation, 2004.

[14] S. Lee, W. Trench, and A. Willis. "Willis's elements of quantity surveying.", John Wiley & Sons; 2011 Mar 1.

[15] R. A. Rashid, M. Mustapa, and S. N. A. Wahid. "Bills of Quantities-Are they still useful and relevant today." In International conference on construction industry, 2006.

[16] W. A. W. A. Bakar, M. A. Jalil, M. Man, Z. Abdullah, and F. Mohd. "Postdiffset: an Eclat-like algorithm for frequent itemset mining." *International Journal of Engineering & Technology* 7, no. 2.28, 2018, pp. 197-199.

[17] W. A. W. A. Bakar, M. Man, M. Man, and Z. Abdullah. "i-Eclat: performance enhancement of Eclat via incremental approach in frequent itemset mining," *Telkomnika* 18, no. 1, 2020, pp. 562-570.

Efficient Function Integration and a Case Study with Gompertz Functions for Covid-19 Waves

Oliver Amadeo Vilca-Huayta[1]
Professor
Department of System Engineering
Universidad Nacional del Altiplano
Puno, Perú

Ubaldo Yancachajlla Tito[2]
Ingeniería en Energías Renovables
Universidad Nacional de Juliaca
Juliaca, Perú

***Abstract*—Numerical algorithms are widely used in different applications, therefore, the execution time of the functions involved in numerical algorithms is important, and, in some cases, decisive, for example, in machine learning algorithms. Given a finite set of independent functions A(x), B(x), ..., Z(x) with domains defined by disjoint, consecutive, and not necessarily adjacent intervals, the main goal is to integrate into a single function F(x) = k1×A(x) + k2×B(x) + … + kn×Z(x), where each activation coefficient k, is one if x is in the interval of the respective domain and zero otherwise. The novelty of this work is the presentation and formal demonstration of two general forms of integration of functions in a single function: The first is the mathematical version and the second is the computational version (with the AND function at the bit level), which is characterized by its efficiency. The result is applied in a case study (Peru), where two regression functions were obtained that integrate all the waves of Covid-19, that is, the epidemic curve of the variable global number of deaths/infected per day, the adjustment provided a highly statistically significant measure of correlation, a Pearson's product-moment correlation of 0.96 and 0.98 respectively. Finally, the size of the epidemic was projected for the next 30 days.**

Keywords*—Covid-19; corona virus; function integration; Gompertz model; machine learning

I. Introduction

Numerical algorithms are important in different applications, they are made up of loops/iterations that contain functions, for example, the cost function in machine learning algorithms.

On the other hand, on some occasions, the integration of functions is necessary, that is, the union of independent functions $G_1(x)$, $G_2(x)$, …, $G_n(x)$, each one in different domains, that is:

$$\begin{aligned} &G_1(x)\,, \quad if\ x \in [x_1, x_1'] \\ &G_2(x)\,, \quad if\ x \in [x_2, x_2'] \\ &\quad\ \ldots \\ &G_n(x)\,, \quad if\ x \in [x_n, x_n'] \end{aligned}$$

Where the function $G_1(x)$ is defined if x is in the interval $[x_1, x_1']$, $G_2(x)$ if x is in the interval $[x_2, x_2']$, …, $G_n(x)$ if x is in $[x_n, x_n']$. Note: Domains are defined by disjoint, consecutive, and not necessarily adjacent intervals:

$$[x_1, x_1'] < [x_2, x_2'] < \cdots < [x_n, x_n']$$

Therefore, the objective is to present new procedures for integrating functions into a single function $F(x) = k_1 \times G_1(x) + k_2 \times G_2(x) + \ldots + k_n \times G_n(x)$, where k_i, with $1 \le i \le n$, is the activation coefficient, that is, it is one if x is in $[x_i, x_i']$ and zero otherwise.

What follows is to apply the results in a case study, that is, the integration in a single function of the different functions that represent the waves of the coronavirus disease (Covid-19). This emergency situation has made it a very important research topic in the entire scientific community [1]–[3].

The Covid-19 has caused deaths and infections since it began. The countries are going through the third and fourth waves and it is not known if others are coming, so it is necessary to build a general regression function for an unlimited number of waves.

The Gompertz model represents sigmoidal behaviour and is suitable for representing the spread of Covid-19. Epidemiologists, biologists, and others use this model for its advantages. There is a detailed review of the Gompertz model in [4]. Then, the study focuses on studying and understanding the global number of deaths/confirmed accumulated, applying the Gompertz model for several waves, that is, an integrated regression function with Gompertz functions ($G_1(x)$, $G_2(x)$, …) for each wave and the prediction of future behaviours. Predictions are important for decision-making in the political, economic, and other fields [5].

Artificial intelligence and its machine learning methods have been applied in different areas and better results have been obtained than traditional methods such as the traditional regression with the normal equation [6]–[11]. In our work, a program has been developed to carry out regression using machine learning, as a case study the country of Peru was selected, which has been one of the countries with the highest mortality rate per inhabitant.

The main part of the research is made up of theoretical results, which are permanent, because mathematical demonstration is used. But, in the application part, which has the purpose of highlighting the importance and usefulness of the results, it is limited to a case study (Peru).

Finally, the results obtained can be applied: In different functions without restriction (for example, in epidemiological functions: Logistics, Bertalanffy, Boltzmann, etc. or a combination thereof), in any data set (for example, Covid-19

data from other countries), in an unlimited number of functions (e.g., in various waves of Covid-19) and in general in all applications that require the integration of functions.

The rest of this research is organized as follows: Section II reviews the related work. Section III explains the methodology. Section IV describes the principal results, while Section V discusses the case study, and applications of the proposed results. Lastly Section VI summarizes the main conclusions of this work.

II. Related Work

Aferni et al. used a basic way of integrating two functions for two consecutive waves of Covid-19, the authors do not present a general way to integrate functions; in addition, it is not possible to generalize the result for more than two functions. In the integrated function F(x), values of zero and one are used for p and the sigmoidal-Boltzmann mathematical model was applied to study the Covid-19 spread in 15 different countries [4].

As far as we know, there is no other work related to the integration of functions. On the other hand, there are several research works regarding the spread of Covid-19 [12]–[14].

III. Methodology

Mathematical proof is the primary form of justification for mathematical knowledge [15]. It is a formal and rigorous method, its validity is permanent, that is, it remains forever. It has no margin of error and is not subject to the assumptions of statistical methods. For this reason, it was used in the first and main part of the investigation.

In order to highlight the importance and usefulness of the theoretical results, in the second part, a regression function composed of several Gompertz functions was built, specifically, for the Covid-19 data (Peru). Linear regression was used with a correlational hypothesis. Linear regression is still a useful and widely used statistical method [16].

A. Data

The data used was the global number of cumulative deaths/infections, which is freely downloadable from the Johns Hopkins University Resource Center repositories from day one to June 27, 2022.

B. Inflection Points and Second Derivative

Let $f(x)$ be a function, which is continuous at a point x_1, $f(x)$ can have a finite or infinite derivative at that point. If, when passing through x_1, the function changes the direction of convexity, then x_1 is called a point of inflection [17].

Second derivative [17], [18] of a function $\boldsymbol{f(x)}$. If x_1 is a point of inflection, and the function has a second derivative in some neighborhood of x_1, which is continuous at the point x_1 itself, then, $\boldsymbol{f''(x) = 0}$.

For example, in the case of Italy, the start of the second wave (inflection point) was calculated with the following procedure: First, a third degree polynomial regression was performed between day 75 and day 250 because it is the interval where the change point is found (in Fig. 1 the regression function is illustrated in blue), second, the inflection point is calculated using the second derivative, finally a red vertical line was drawn to highlight the start day of the second wave.

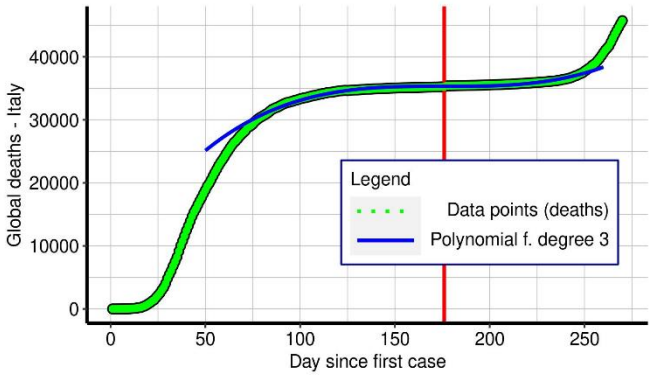


Fig. 1. Global Death by Days (Italy), Polynomial Function Regression of Degree 3 (Blue) and Inflection Point (Red Line).

The cubic regression function is given by (R statistical software was used):

$$f(x) = 7446.998 + 476.22x - 2.710676x^2 + 0.005140938x^3$$

The inflection point was obtained by calculating the second derivative and solving the equation $f''(x) = 0$, it is given by (the free Maxima software was used):

$$x = 175.7575498219713$$

Then, nonlinear regression can be performed using the Gompertz function or another for each wave.

C. The Gompertz Model

The Gompertz curve/function/model is a function for a time series, named after Benjamin Gompertz (1779–1865), who was born in the City of London) [19]–[22]. It is a special case of the generalized logistic function. In [23] the Gompertz function is described, classified, and explained, it has different variants, the best known is the following:

$$G(x) = a \cdot e^{-e^{b-c \cdot x}}$$

Also represented with the exp function as follows:

$$G(x) = a * \exp(-\exp(b - c * x))$$

Where, G(x) is the expected value (e.g., deaths) as a function of time x (for example days since the first case), a is the upper asymptote, b sets the displacement along the x-axis (translates the graph to the left or right), c is the growth-rate coefficient (which affects the slope), e is Euler's Number (e = 2.718281828459045), and exp(x) is e^x.

D. Fit Assessment Measures

The evaluation measures, known as goodness of fit measures that will be used in the research are the Pearson correlation coefficient (R) and the determination coefficient (R^2).

$$R^2 = 1 - \frac{MSE}{MST}$$

IV. RESULTS

To integrate several Gompertz functions (one for each wave) in a single formula, first the inflection points are calculated, then the regression is carried out to obtain the Gompertz (or other) functions and it ends up integrating the functions in a single formula. With this result, forecasts can be made.

A. Calculation of Inflection Points

It is achieved by performing the cubic linear regression (third degree polynomial) in the respective intervals, then using the second derivative of these functions, the inflection points are obtained, which can be interpreted as the day on which a wave begins (after a wave as seen in Fig. 1 and Fig. 2). In Peru, the first inflection point is day 253 and second is day 596.

B. Integration of Functions

Let's start with the basic case of the union of two functions that represent two successive waves (of deaths or another accumulated variable), for illustrative purposes the Gompertz function will be used, although the results of this section are general, that is, can be applied to other functions (e.g., the Boltzmann function).

Let $G_1(x)$ y $G_2(x)$, two Gompertz functions (without losing generality), which correspond to two successive waves, be integrated into a single function by adding a characteristic/variable in the database called p, which indicates with a single distinctive number all the rows (days) that correspond to a particular wave. The first wave is assigned a number (e.g., one), the second wave is marked with another number (e.g., two).

The data is structured as shown in Table I. In the case of Italy (as also seen in Fig. 1), from day 1 to day 175 belong to the first wave and from day 176 onward to the second wave. (Basic study for Italy that does not include integrated waves is found in [24] and [25]).

TABLE I. DATABASE OF THE GLOBAL NUMBER OF DEATHS ACCORDING TO DAY - ITALY, THE LAST CHARACTERISTIC (COLUMN) ON THE RIGHT INDICATES TO WHICH FUNCTION BELONGS

Day	***Date***	***Deaths global (Accumulated)***	***p***
1	2/21/20	1	1
2	2/22/20	2	1
...	...	...	...
176	8/14/20	35234	2
177	8/15/20	35392	2
...	...	...	...

Then, the integrated regression function is given by:

$$F(x) = (2-p) * G_1(x) + (p-1) * G_2(x)$$

In this way, if in F(x), p is replaced with 1, $G_1(x)$ is activated and $G_2(x)$ is eliminated, conversely, if p is 2, $G_2(x)$ is activated. The function F(x) can be simplified if values of zero and one are used for p as was done in [4], however, it is not useful for generalizing over n functions.

General case mathematical version, if several functions are considered, whose quantity is specified with the value of n, proceed as follows:

Let $G_1(x)$, $G_2(x)$, … , $G_n(x)$ be functions, which correspond to n successive waves, are integrated into a single function, controlling the activation through a coefficient C_p as follows: $F(x) = C_1 \times G_1(x) + C_2 \times G_2(x) + \ldots + C_n \times G_n(x)$ adding a feature (variable p) to the training samples, as the basic case (e.g. successive numbers 1, 2, 3, ... , p).

A table of coefficients is constructed to discriminate functions. Then, it is generalized with a new general formula for the coefficients.

TABLE II. COEFFICIENT TABLE FOR THE SUCCESSIVE FUNCTIONS

p	**$G_1(x)$**	**$G_2(x)$**	**$G_3(x)$**	**$G_4(x)$**	**...**	**$G_n(x)$**
1		1-p	1-p	1-p	...	1-p
2	2-p		2-p	2-p	...	2-p
3	3-p	3-p		3-p	...	3-p
4	4-p	4-p	4-p		...	4-p
...	...	...	...	...	...	...
n	n-p	n-p	n-p	n-p	...	

Then, the formula for the coefficients of the sequence of functions according to the parameter p, is given by:

$$C_p = \sum_{p=1}^{n} \frac{\prod_{i=1}^{p-1}(i-p)\prod_{j=p+1}^{n}(j-p)}{(\mathrm{n}-\mathrm{p})!\,(-1)^{p-1}(\mathrm{p}-1)!}$$

Where, p is the number of the function or wave, C_p is the coefficient for the function $G_p(x)$, n is the total number of functions to integrate.

Proof. Let the function be Gp(x), its coefficient is calculated as follows, the multiplication of (n-1) factors is required except the one corresponding to row p (to activate the function in question), the factors are described like column Gp(x) in Table II, it is achieved with two products $\prod_{i=1}^{p-1}(i-p)$ and $\prod_{j=p+1}^{n}(j-p)$ for factors above and below p respectively. Replacing p in the factors generates two factorials that offset by dividing (p-1)! and (n-p)! generated by the product of the top and bottom numbers. Finally, divide by $(-1)^{p-1}$ to nullify the negative result that occurs when p is even.■

Therefore, the final function is given by:

$$F(x) = \sum_{p=1}^{n} \frac{\prod_{i=1}^{p-1}(i-p)\prod_{j=p+1}^{n}(j-p)}{(\mathrm{n}-\mathrm{p})!\,(-1)^{p-1}(\mathrm{p}-1)!} G_p(x)$$

The advantage of this formula (mathematical version) is that it does not depend on a programming language, compiler, or binary representation of the numbers, but the disadvantage is the execution time required to calculate each coefficient, which is $\Theta(n)$, where n is the number of waves or functions,

specifically among others requires (n+1) multiplications and one division.

For example, if you have three waves, the coefficient for the first function $G_1(x)$ is given by:

$$F(x) = \frac{(2-\mathrm{p})(3-\mathrm{p})}{2}\mathrm{G}_1(\mathrm{x}) + \cdots$$

General case, the computational version (with function at the bit level). It is possible to perform the integration by a simpler method, that is, using bitwise operations, specifically using the bitwise AND (&) operator. Unlike the previous method, the p values for each group (function) must be recorded in powers of two starting from one, that is, 1, 2, 4, 8, 16, ..., 2^{n-1}:

$$F(x) = (1\&p)G_1(x) + \frac{(2\&p)}{2}G_2(x) + \frac{(4\&p)}{4}G_3(x) +$$

$$\ldots + \frac{(2^{n-1}\&p)}{2^{n-1}}G_n(x)$$

Where, the & operator represents the bitwise AND operator.

Proof. Given the conditions, the result of $2^{n-1}\&p$ is equal to 2^{n-1}, if and only if both operands have the same value (by the definition of the AND operation). Then only one division is required to get the 1, which finally selects the function.■

The method is simple and efficient, the execution time to calculate each coefficient is constant, but it requires the AND function at the bit level. In the statistical software R, it has the bitwAnd(a,b) function, which, according to the notation considered, calculates a&b. An empirical comparison of the performance of the two methods is not necessary, the difference is obvious.

Finally, source code is presented for the statistical program R that represents three integrated Gompertz functions. The method can be implemented in any other programming language (e.g., Python).

```
F = function ( x , p ) {
   bitwAnd(1,p)    * (ga*e^(-e^(gb-gc*x)) ) +
   bitwAnd(2,p)/2 * (ga2*e^(-e^(gb2-gc2*(x-a2) )) +dif1 ) +
   bitwAnd(4,p)/4 * (ga3*e^(-e^(gb3-gc3*(x-b2) )) +dif2 )
}
```

V. The Case Study

A. Case Study Analysis of the Global Number of Deaths from Covid-19 in Peru

To apply the procedure described in this work, the regression function of the global number of deaths from Covid-19 in Peru was analysed using three Gompertz functions for each wave. Day one corresponds to the first case of death that occurred on March 3, 2020, and the time series extends until June 27, 2022 (which makes up a total of 844 days). Fig. 2 shows the observed data (in green), the Gompertz1 function for the first wave (blue), the Gompertz2 function for the second wave (red), and the Gompertz3 function for the third wave (black).

To calculate the day on which the first wave ends and the day on which the second wave begins (with a growth in the number of deaths), a third-degree simple polynomial regression was carried out, to this formula the second derivative was applied as shown described in the methods section (for the purpose of finding the inflection point). On day 253 the first wave ends and on day 254 the second wave begins. The same procedure was carried out to calculate the second inflection point.

a) Regression Function: The formula of the regression function is:

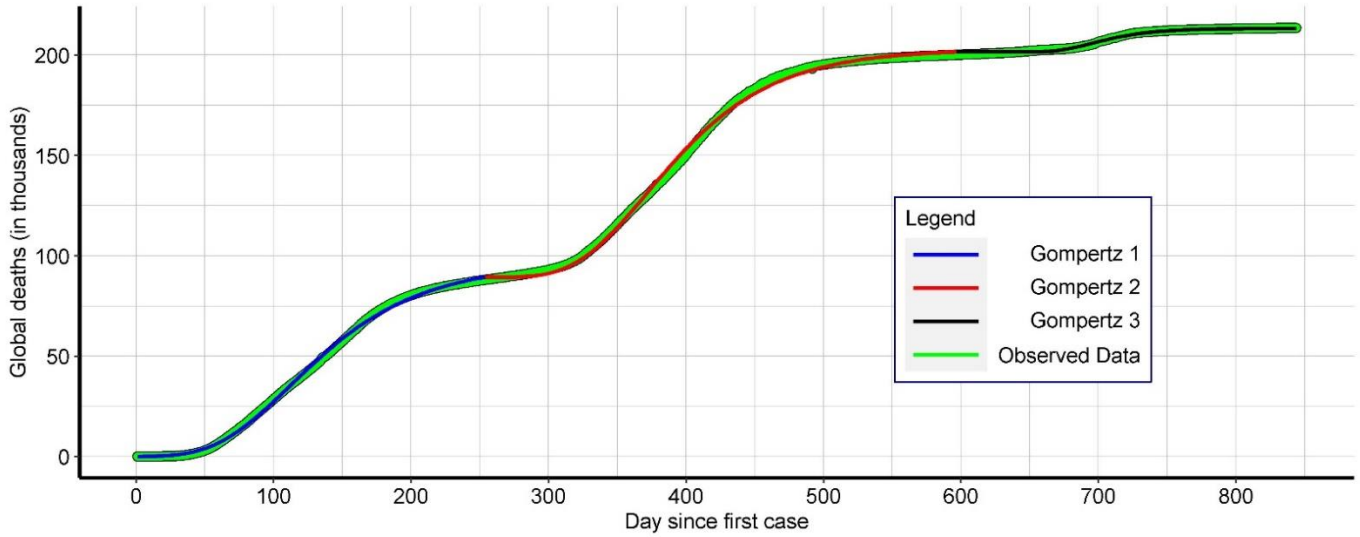


Fig. 2. Number of Global Death (in Thousands) by Day and Gompertz Regression Function - Perú (2020-2022).

$$F(x,p) = (1\&p)95.82221e^{-e^{2.095506-0.01870962x}} + \frac{(2\&p)}{2}\left(113.7772e^{-e^{2.307933-0.01948187(x-253)}} + 89.20994\right) + \frac{(4\&p)}{4}\left(11.72726e^{-e^{4.016593-0.04029681(x-596)}} + 201.5629\right)$$

This function is the one observed in Fig. 2, each wave with a different colour. The Gompertz model adjusted to the series of the accumulated number of deceased, reports a Pearson correlation coefficient R = 0.9577994 and an explained variance of 91.73797%, quite acceptable measurements of the adjustment made. The alternative hypothesis is accepted: the correlation is not equal to zero (t = 96.691, df = 842, p-value < 2.2e-16).

b) Prediction: The prediction was made for 30 days, in Fig. 3 you can see the projection that corresponds to days greater than 844 (green).

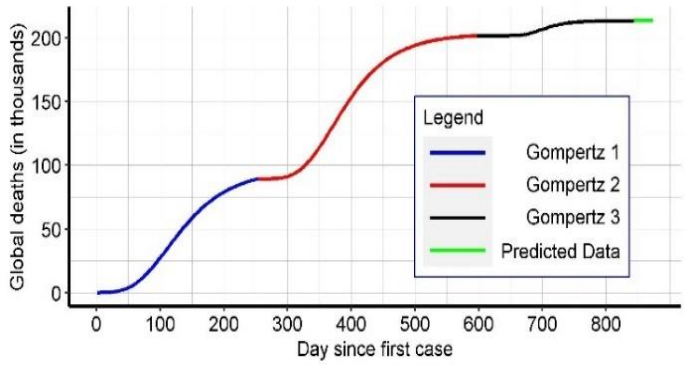


Fig. 3. Prediction of the Number of Global Death (in Thousands) by Day - Perú (2020-2022).

In Fig. 3, and particularly in the predicted data, a small increase is observed, so it is necessary to take measures.

B. Analysis of the Global Infected Number - Peru

The regression function of the global infected number (Covid-19 in Peru) was analysed using three Gompertz functions for each wave. Day one corresponds to the first infected case (March 3, 2020), and the time series extends until June 27, 2022 (which makes up a total of 844 days). Fig. 4 shows the observed data (in green), the Gompertz1 function for the first wave (blue), the Gompertz2 function for the second wave (red), and the Gompertz3 function for the third wave (black).

a) Regression Function: This function is the one observed in Fig. 4, each wave with a different colour. The Gompertz model adjusted to the series of the accumulated number of infected, reports a Pearson's product-moment correlation R = 0.9814708 and an explained variance of 91.73797%, quite acceptable measurements of the adjustment made. The alternative hypothesis is accepted: the correlation is not equal to zero (t = 148.63, df = 842, p-value < 2.2e-16).

$$F(x,p) = (1\&p)1263.777e^{-e^{1.892196-0.01259441x}} + \frac{(2\&p)}{2}1262.33e^{-e^{2.092404-0.01745463x}} + \frac{(3\&p)}{3}1402.889e^{-e^{11.10519-0.08365515x}}$$

In this case of the number of infections, no predictions are made because the last Gompertz curve G3(x) does not fit adequately to obtain consistent results. Finally, it can be said that the function integration method worked well in both cases (deaths and confirmed).

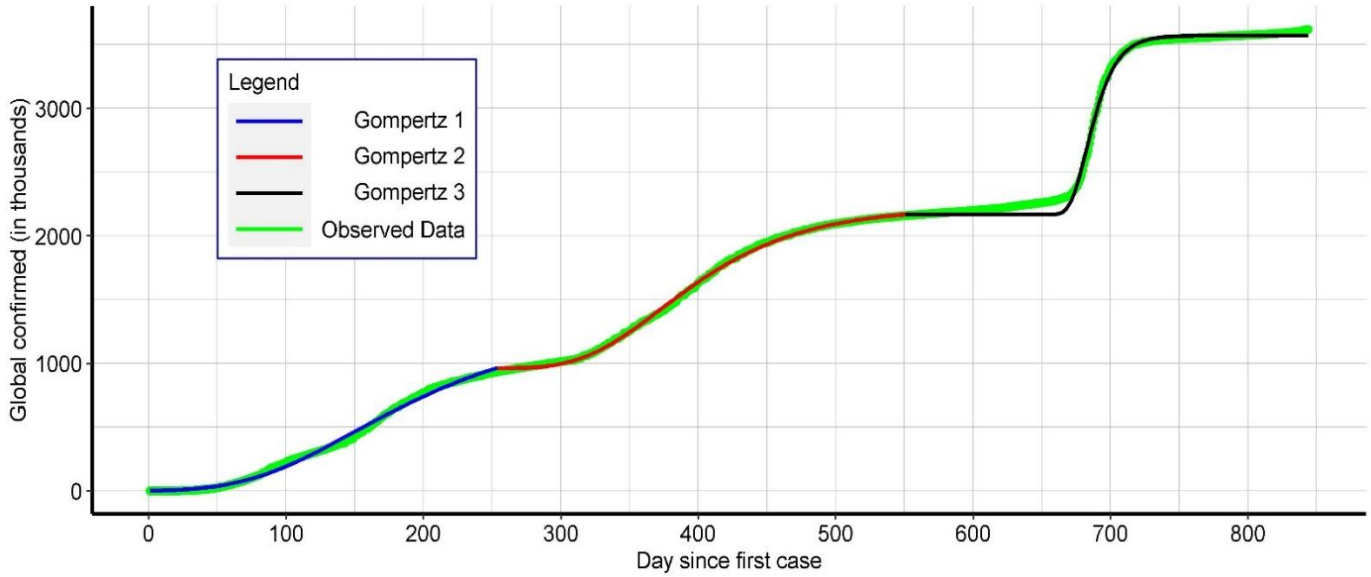


Fig. 4. Number of Global Infected (in Thousands) by Day and Gompertz Regression Function - Perú (2020-2022).

VI. Conclusion

The contribution of this research is the presentation and illustration of two ways of integrating *n* regression functions (which may correspond to n waves of covid-19 or others). The first is the mathematical version, independent of devices such as the binary representation in the computer, and the second one is the computational version that has the advantage of being simple and efficient in time, specifically, in the calculation of the coefficients (with a constant time complexity). These results are general in relation to the cited literature and have many applications.

The epidemic curve of the number of deaths/infections was obtained with three Gompertz models integrated into one function, the adjustment provided a correlation measure that is statistically quite reliable, so forecasts were obtained for 30 days, that is, for the month of July 2022, it is concluded that the model fits the data well and is good for forecasting. And, given the slight outbreak, it is necessary to follow preventive measures to prevent the spread of Covid-19 (with specific emphasis on Peru).

Finally, the detailed explanation and interpretation of the linear regression function and the Gompertz functions that compose it, are useful to describe and compare the waves of Covid-19, however, they go beyond the objectives of the research.

VII. Funding

The research has not received funds for this work.

VIII. Consent

Informed consent was obtained from all individual participants included in the study.

IX. Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgment

The authors thank Johns Hopkins University for freely organizing and providing the Covid-19 data.

References

[1] A. M. Ćmiel and B. Ćmiel, 'A simple method to describe the COVID-19 trajectory and dynamics in any country based on Johnson cumulative density function fitting', *Scientific Reports 2021 11:1*, vol. 11, no. 1, pp. 1–10, Sep. 2021, doi: 10.1038/s41598-021-97285-5.

[2] G. Ramírez-Valverde, B. Ramírez-Valverde, G. Ramírez-Valverde, and B. Ramírez-Valverde, 'Modelo estadístico para defunciones y casos positivos de COVID-19 en México', *EconoQuantum*, vol. 18, no. 1, pp. 1–20, Dec. 2021, doi: 10.18381/EQ.V18I1.7223.

[3] N. P. Dharani and P. Bojja, 'Analysis and Prediction of COVID-19 by using Recurrent LSTM Neural Network Model in Machine Learning', *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, p. 2022, 2022, doi: 10.14569/IJACSA.2022.0130521.

[4] A. el Aferni, M. Guettari, and T. Tajouri, 'Mathematical model of Boltzmann's sigmoidal equation applicable to the spreading of the coronavirus (Covid-19) waves', *Environmental Science and Pollution Research*, vol. 28, no. 30, pp. 40400–40408, Aug. 2021, doi: 10.1007/S11356-020-11188-Y/FIGURES/8.

[5] G. L. Watson *et al.*, 'Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model', *PLoS Comput Biol*, vol. 17, no. 3, p. e1008837, Mar. 2021, doi: 10.1371/JOURNAL.PCBI.1008837.

[6] R. K. Mishra, S. Urolagin, J. A. A. Jothi, N. Nawaz, and H. Ramkissoon, 'Machine Learning based Forecasting Systems for Worldwide International Tourists Arrival', *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, pp. 55–64, Jan. 2021, doi: 10.14569/IJACSA.2021.0121107.

[7] P. C. Albuquerque, D. O. Cajueiro, and M. D. C. Rossi, 'Machine learning models for forecasting power electricity consumption using a high dimensional dataset', *Expert Syst Appl*, vol. 187, p. 115917, Jan. 2022, doi: 10.1016/J.ESWA.2021.115917.

[8] G. Papacharalampous and H. Tyralis, 'A review of machine learning concepts and methods for addressing challenges in probabilistic hydrological post-processing and forecasting', Jun. 2022, doi: 10.48550/arxiv.2206.08998.

[9] G. C. Pinasco *et al.*, 'An interpretable machine learning model for covid-19 screening', *Journal of Human Growth and Development*, vol. 32, no. 2, pp. 268–274, Jun. 2022, doi: 10.36311/JHGD.V32.13324.

[10] Y. Xiong, Y. Ma, L. Ruan, D. Li, C. Lu, and L. Huang, 'Comparing different machine learning techniques for predicting COVID-19 severity', *Infect Dis Poverty*, vol. 11, no. 1, p. 19, Feb. 2022, doi: 10.1186/S40249-022-00946-4/FIGURES/4.

[11] Y. Alali, F. Harrou, and Y. Sun, 'A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models', *Scientific Reports 2022 12:1*, vol. 12, no. 1, pp. 1–20, Feb. 2022, doi: 10.1038/s41598-022-06218-3.

[12] H. Nieto-Chaupis, 'Modeling and Interpretation of Covid-19 Infections Data at Peru through the Mitchell's Criteria', *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 9, pp. 717–722, 2020, doi: 10.14569/IJACSA.2020.0110986.

[13] N. P. Dharani and P. Bojja, 'Analysis and Prediction of COVID-19 by using Recurrent LSTM Neural Network Model in Machine Learning', *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, p. 2022, 2022, doi: 10.14569/IJACSA.2022.0130521.

[14] M. Torky, M. S. Torky, A. A. Azza, A. E. Hassanein, and W. Said, 'Investigating Epidemic Growth of COVID-19 in Saudi Arabia based on Time Series Models', *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 459–466, 2020, doi: 10.14569/IJACSA.2020.0111256.

[15] Y. Hamami, 'MATHEMATICAL RIGOR AND PROOF', *The Review of Symbolic Logic*, vol. 15, no. 2, pp. 409–449, Jun. 2022, doi: 10.1017/S1755020319000443.

[16] G. James, D. Witten, T. Hastie, and R. Tibshirani, 'Linear Regression', pp. 59–128, 2021, doi: 10.1007/978-1-0716-1418-1_3.

[17] R. Shende, G. Gupta, and S. Macherla, 'Determination of an inflection point for a dosimetric analysis of unflattened beam using the first principle of derivatives by python code programming', *Reports of Practical Oncology and Radiotherapy*, vol. 24, no. 5, pp. 432–442, Sep. 2019, doi: 10.1016/J.RPOR.2019.07.009.

[18] P. Román-Román, J. J. Serrano-Pérez, and F. Torres-Ruiz, 'A Note on Estimation of Multi-Sigmoidal Gompertz Functions with Random Noise', *Mathematics 2019, Vol. 7, Page 541*, vol. 7, no. 6, p. 541, Jun. 2019, doi: 10.3390/MATH7060541.

[19] T. B. L. Kirkwood, 'Deciphering death: a commentary on Gompertz (1825) "On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies"', *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1666, Apr. 2015, doi: 10.1098/RSTB.2014.0379.

[20] R. Kundu, H. Basak, P. K. Singh, A. Ahmadian, M. Ferrara, and R. Sarkar, 'Fuzzy rank-based fusion of CNN models using Gompertz function for screening COVID-19 CT-scans', *Scientific Reports 2021 11:1*, vol. 11, no. 1, pp. 1–12, Jul. 2021, doi: 10.1038/s41598-021-93658-y.

[21] R. A. Conde-Gutiérrez, D. Colorado, and S. L. Hernández-Bautista, 'Comparison of an artificial neural network and Gompertz model for predicting the dynamics of deaths from COVID-19 in México', *Nonlinear Dyn*, vol. 104, no. 4, pp. 4655–4669, Apr. 2021, doi: 10.1007/S11071-021-06471-7/FIGURES/8.

[22] D. Garcia-Vicuña, L. Esparza, and F. Mallor, 'Hospital preparedness

during epidemics using simulation: the case of COVID-19', *Cent Eur J Oper Res*, pp. 1–37, Sep. 2021, doi: 10.1007/S10100-021-00779-W/TABLES/9.

[23] K. M. C. Tjørve and E. Tjørve, 'The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family', *PLoS One*, vol. 12, no. 6, p. e0178691, Jun. 2017, doi: 10.1371/JOURNAL.PONE.0178691.

[24] N. Chintalapudi, G. Battineni, and F. Amenta, 'Second wave of COVID-19 in Italy: Preliminary estimation of reproduction number and cumulative case projections', *Results Phys*, vol. 28, p. 104604, Sep. 2021, doi: 10.1016/J.RINP.2021.104604.

[25] I. Chalkiadakis, H. Yan, G. W. Peters, and P. v. Shevchenko, 'Infection rate models for COVID-19: Model risk and public health news sentiment exposure adjustments', *PLoS One*, vol. 16, no. 6, pp. 1–39, Jun. 2021, doi: 10.1371/JOURNAL.PONE.0253381.

Enhancement of Low-Light Image using Homomorphic Filtering, Unsharp Masking, and Gamma Correction

Tan Wan Yin[1], Kasthuri A/P Subaramaniam[2], Abdul Samad Bin Shibghatullah[3]
Institute of Computer Science and Digital Innovation
UCSI University
Kuala Lumpur, Malaysia

Nur Farraliza Mansor[4]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Besut
Terengganu, Malaysia

***Abstract*—Now-a-days, a digital image can be found almost everywhere, and digital image processing plays a huge role in analyzing and enhancing the image so that it can be delivered in a good condition. Color distortion and loss of image details are the common problems that were faced by low-light image enhancement methods. This paper introduces a low-light image enhancement method that applied the concept of homomorphic filtering, unsharp masking, and gamma correction. The aim of the proposed method is to minimize the two problems stated while producing images of better quality when compared to the other low-light image enhancement methods. An objective evaluation was done on the proposed method, comparing the results produced by the enhanced method with other two existing low-light image enhancement methods. The results obtained showed the proposed method outshines the other two existing low-light image enhancement method in maintaining the image details and producing a natural looking image, achieving the lowest Mean Square Error (MSE) and Lightness Order Error (LOE) scores, and has the highest Features Similarity Index color (FSIMc), Features Similarity Index (FSIM), Structure Similarity Index (SSIM), and Visual information fidelity (VIF) scores. Future studies that should be made on this research are to implement dehaze and denoise functionality into the low-light image as well as enabling it to be applicable in real-time scenarios.**

Keywords*—Low-light image; gamma correction; homomorphic filtering. low-light enhancement; unsharp masking

I. INTRODUCTION

It is undeniable that digital image processing is important as it helps to enhance the quality of an original image. A low-light image, as its name suggests, is an image that is captured in a low-light environment. This type of image is typically found when the image is captured in the nighttime. Images captured during nighttime are lacking because the amount of light captured in the image is low. This would result in low visibility and details in objects that are captured and cause color distortion in the image.

Low-light image enhancement methods help to boost the features of objects captured in an environment where the light source is minimal. In a situation where hit-and-run occurred during nighttime and the car was captured in a CCTV, low light image enhancement methods can be used to obtain the visual information of image like the color of car, the shirt color of the driver, etc. and followed by digital image processing to gain more information on the accident occurs, for example, getting the car plate number to track the irresponsible driver. This becomes the motivation for the proposed low light image enhancement. To produce a clear and bright image, low-light image enhancement method like Retinex method, histogram equalization, neural network, gamma correction and homomorphic filtering were introduced [1]. A significant amount of research has been made on the methods of enhancing low-light image enhancement in recent years, however, there is still room for improvements in the techniques used to enhance low-light images. The problems faced in the enhancement of low-light images include color distortion and loss of image detail. Colors are hard to be distinguished in a low-light environment, thus, it is important to enhance low-light image while retaining the color of the image. Loss of image details is another problem faced during the enhancement process.

In this paper, homomorphic filtering, unsharp masking, and gamma correction techniques have been applied to create a low-light image enhancement method. The input image will be processed with homomorphic filtering where the parameters values of the Gaussian high pass filter available for customization have been applied and Fast Fourier Transform (FFT) is used so that the time taken for the process can be cut down. Subsequently, to sharpen the image, unsharp masking is applied to the image. Finally, the requirement for gamma correction of the image is defined into four different states, low, medium, high, and none depending on the luminance. The gamma will be influenced by the luminance of the image and different values will be set as the gamma according to the different states of the image. The contribution of the research are as follows:

- Introduced a method to enhance low-light images by classifying the state of low-light images into three stages
- Enhanced low-light images with minimal color distortion issues
- Preserved the image details and produce a more natural enhanced image

- Produced a better-quality image compared to the other two low-light image enhancement methods.

There are a total of eight sections in this paper. The first section introduces the existing low light image enhancement methods and the related works, as mentioned above. Section II describes the related works. Section III mentioned about the image enhancement techniques that were applied in the proposed method. The proposed method was explained in detail in Section IV followed by experimental results that are shown in Section V. Discussions were presented in Section V, and the final section would be concluding this research.

II. Related Work

Low-light image enhancement methods, namely Retinex method, histogram equalization, and neural network were some of the image enhancement methods that can be used to enhance a low-light image.

A. Retinex Method

The Retinex method uses the concept of human visual systems which perform automatic color and brightness adjustments on scenery that are captured by the human eyes [2]. This method expresses the image using the illumination and the reflection of the image [3]. Retinex methods include a Single-scale Retinex (SSR), multi-scale Retinex (MSR), and multi-scale Retinex with color restoration (MSRCR). SSR as well as MSR algorithms [4],[5] are algorithms that apply the Gaussian surround function on the input image. Both methods are used to obtain the reflection image by estimating the illumination level, the formula of SSR and MSR are listed below:

$$\text{SSR} = \log R_i(x,y) = \log I_i(x,y) - \log[\, Ke^{\left(-\frac{x^2+y^2}{\sigma^2}\right)} * I_i(x,y)] \quad (1)$$

Where $I(x, y)$ is the input image, $R(x, y)$ is the reflection image, i is the RGB color channels, (x, y) is the position of pixels in the image, K is the normalization factor while e is the exponential function, σ is the scale parameter, and * is convolution operator.

$$\text{MSR} = \log R_i(x,y) = \sum_{k=1}^{N} \omega_k \{\log I_i(x,y) - \log\left[Ke^{\left(-\frac{x^2+y^2}{\sigma^2}\right)} * I_i(x,y)\right]\} \quad (2)$$

Here, k is Gaussian surround scales, the number of scales is represented with N, and ω are the scale weights. MSR has the advantage of having multiple scales which further enhances the details of image and contrast and produces images with improved visual effect.

An issue of color distortion effect might occur with SSR and MSR methods which leads to the introduction of MSRCR method [6], [7]. In this method, the color recovery of each color channel, C will be calculated based on the proportional relationship between RGB channels of the raw image and used to overcome the color distortion problem, which is:

$$C_i(x,y) = \beta \times \log(\alpha \times \frac{I_i(x,y)}{\sum_i^3 I_i(x,y)}) \quad (3)$$

By combining color recovery of each color channel with MSR, the equation of MSRCR will be formed, where:

$$\text{MSRCR} = \sum_{k=1}^{N} C_i\, \omega_k \{\log I_i(x,y) - \log\left[Ke^{\left(-\frac{x^2+y^2}{\sigma^2}\right)} * I_i(x,y)\right]\} \quad (4)$$

Although MSRCR successfully solved the color distortion problem, it would lose image details in the bright region.

Retinex-based method is widely applied in low-light images, for instance, a fast algorithm based on Retinex was proposed by Liu et al. [8], where the low-light image will be converted to HSV color space, and linear function is used to stretch the gray level in V component followed by Retinex model which is applied to enhance the brightness of a low-light image. The method solved the problem of uneven brightness and make the low-light area clearer, yet there is a problem where this method will cause the details of the brighter area in the input image become vague due to the brightness enhancement. In [9], a Retinex model was able to produce a result where brightness and contrast were improved while preserving the details of the image as well as suppressed noise interference through performing various processing in illumination image estimation, reflection image acquisition, and post-processing. However, this method is time-consuming and causes noise amplification problems when input images have a higher amount of light loss. The low-light image enhancement technique by Shi et al. [10] also used the Retinex method and mixed with a generative adversarial network (GAN) to enhance an image under low-light conditions, although the method caused problems like noisy and overly enhanced results in low-light images, this method proved to be very useful in low-light images with very minimal light where the details of these type of image can be seen clearly. Another application of Retinex is observed in [11], where Retinex is applied in the low-light image for the purpose of autonomous vehicles, and successfully enhances the image illumination and improves the detection of the vehicle yet the problem of time consumed for the algorithm must be considered as the method should be working in real-time. In short, the Retinex method is useful in handling color distortion issues and sharpening capability but since Gaussian filtering is applied in the Retinex method, it can be a very complex, and the sharp boundary of an image might cause the image to be too bright.

B. Histogram Equalization

Histogram equalization (HE) is an image enhancement technique where in the grey level of the image, the smaller pixel population would be compressed whereas the larger pixel population is stretched to occupy a wide range [12]. The grey level of the image is then equalized. This method is widely used to improve image contrast [13] but it can cause over-enhanced noise and loss of edges of objects [14]. Using the principles of histogram equalization, the image can be described as:

$$I = \frac{n_{gL}}{Tp},\ (gL = 0, 1, 2, \dots, L-1) \quad (5)$$

Where I is the probability of grey level in the image, Tp is the total pixels in the image, gL is the grey level, and ngL is the number of pixels in grey level. The cumulative distribution function (CDF) of the grey level of the image I can be evaluated as:

$$\text{CDF} = \sum_{r=0}^{gL} I(r), \ (gL = 0, 1, 2, \dots, L-1) \quad (6)$$

Generally, histogram equalization (HE) performs equalized grey level distribution on the original image based on CDF to produce an enhanced image. This method can be executed in real-time, but it has the drawback of changing the brightness of the image. To tackle the problem faced in HE, mean preserving Bi-histogram equalization (BBHE) has been introduced which successfully preserved some of the original brightness of the image. However, it requires greater brightness preservation, which leads to the creation of the Minimum Mean Brightness Error Bi-Histogram Equalization (MMBEBHE) method [15]. This method successfully preserved more brightness and avoided excessive enhancement of the image, however, the Absolute Mean Brightness Error (AMBE) of every possible threshold level needed to be calculated through a full BBHE process would need a large computation process. Contrast Limited Adaptive Histogram Equalization (CLAHE) is another type of histogram equalization method introduced by Reza [16]. Through the algorithm, block effects in the enhancement process are weakened and enhancement of contrast has been limited using a threshold, which avoids overly enhanced contrast however one drawback observed is that the clip limit has to be manually entered. Another histogram equalization method was proposed by Zhuang and Guan [17], namely Mean and Variance based Sub image Histogram Equalization (MVSIHE) which enhances contrast and preserves details of the image; however, the methods are only tested in black and white images. Another method that used adaptive histogram equalization to enhance the contrast of probability distribution function was introduced by Sirajuddeen et al. [18] and able to enhance the image while preserving the details of an image, yet the method can cause oversaturation of color. Although HE can effectively improve contrast and image details, there are several drawbacks which include easily generated noise, color distortion, and image distortion.

C. Neural Network

A neural network is another method that is used to enhance and produce a clearer image with better quality. It is a method based on machine learning and since it is data-driven, the data collected are very important as they will affect the result produced. There are various types of image enhancement methods which apply the neural network, among them, deep convolutional neural network (CNN) technology has received a lot of attention [19]. Besides using a neural network to enhance a low-light image, it can also be used in denoising an image [20].

A low-light image enhancement method applying a neural network is introduced by Gómez et al. [21] which focuses on using CNN to improve low-light images that are captured from high-speed video endoscopy. They use a deep learning approach to enhancing the medical image that is targeted toward patients with laryngeal disorders, however, they lack ground-truth images. Thus, they introduce a method to generate darkened and realistic training images, and results show that the method proposed outperforms existing enhancement methods on the medical images. Besides, Ha et al. [22] applied CNN in CIELAB color space to enhance the low-light image. Their main idea is to split the low-light image into luminance and chrominance components before applying image enhancement on the respective components. The method proved to be very useful in removing undesired artifacts and preventing color distortion. Zero-Reference Deep Curve Estimation (Zero-DCE) is another low-light image enhancement method using a neural network that is proposed by Guo et al. [23] in which a set of best-fitting Light-Enhancement curves (LE-curves) would be estimated for the image using a Deep Curve Estimation Network (DCE-Net) and the enhanced image would be obtained after the curves are applied iteratively in the pixels of an image. The zero-DCE method can generate an image that fits various lightning conditions and performs well although there is a lack of reference images, however, the problem of the generated noises should be tackled in this method. A pipeline neural network is introduced by Guo et al. [24] which consists of a low-light image enhancement net (LLIE-net) and denoising for enhancement of the low-light image. They proved that MSR can be considered as CNN and using wavelet transformation to improve MSR results, thus, an end-to-end convolution network is proposed to enhance the low-light image. The method introduced can produce a quality image without needing to adjust the parameters manually but show lacking in some image as there is a limited dataset. The main drawback of using a neural network would include needing a lot of datasets to produce a well-enhanced image.

III. Image Enhancement Techniques

Homomorphic filtering and gamma correction are popular techniques that are used to enhance low-light image [25]. This section briefly introduces the two techniques.

A. Homomorphic Filtering

Homomorphic filtering is one of the techniques used in enhancing an image in the frequency domain [26]. Homomorphic filtering would convert the image to the frequency domain and enhance the image using Fourier transform and apply a high pass filter before converting the image back to the spatial domain [27].

Fig. 1 shows the homomorphic filtering process flowchart. Here, $F(x, y)$ is the input image whereas $G(x, y)$ is the enhanced image. Log represents logarithmic transform that will be applied on the input image, FFT is applied on both illumination and reflection component, then H (u, v) is frequency filtering function will be applied, the output is then inversed using IFFT, and applied Exp, that is exponential before producing the enhanced image. The image $I(x, y)$ can be represented as a multiplication of illumination, $L(x, y)$ and reflectance of an image, $R(x, y)$:

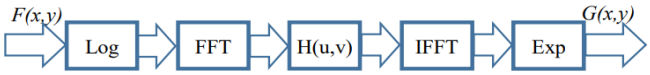


Fig. 1. Homomorphic Filtering Process Flowchart [1].

$$I(x, y) = L(x, y) \times R(x, y) \quad (7)$$

B. Gamma Correction

The gamma correction enhances the low-light image by adjusting the contrast of an image [28], enhancing the pixel intensity of the low-light area. The gamma correction method is also one of the fast and efficient ways to enhance a low-light image [29]. The general formula for this transformation is [30]:

$$g(x, y) = f(x, y)^{\gamma} \quad (8)$$

In this formula, γ is the gamma correction parameter, Fig. 2 shows some examples of gamma function transformation.

The output will be linear if γ = 1, which would return the same image. When γ>1, the low grey value area will be stretched, and the high grey value area will be compressed. In contrast, when γ<1, the low grey value will be compressed, and the high grey value will be stretched [31].

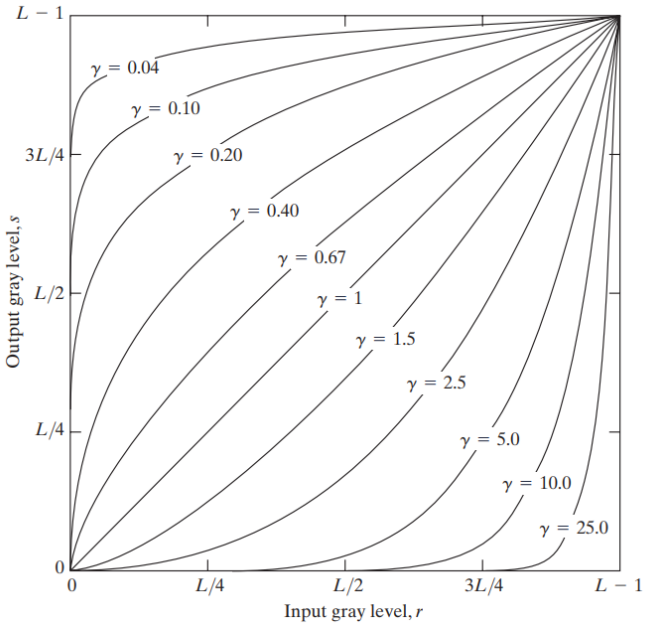


Fig. 2. Gamma Correction [30].

IV. The Proposed Method

A digital image can be expressed as $f(x, y)$, where all values in the function are discrete quantities and finite. The notation of coordinates, where the image has P rows and Q columns originating from point $f(0,0)$ is shown below [32].

$$f(x, y) = \begin{bmatrix} f(0,0) & f(1,0) & \dots & f(0, Q-1) \\ f(1,0) & f(1,1) & \dots & f(1, Q-1) \\ \vdots & \vdots & \ddots & \vdots \\ f(P-1, 0) & f(P-1,1) & \cdots & f(P-1, Q-1) \end{bmatrix} \quad (9)$$

To solve the problem of color distortion and loss of image details, this paper proposes a low-light image enhancement method using a homomorphic filtering method and gamma correction. Following are the steps carried out using the proposed algorithm:

1) Homomorphic filtering is applied on the RGB image
2) Unsharp masking is done on the enhanced image
3) Gamma correction based on the luminance of image is applied to produce the enhanced image.

Fig. 3 shows the algorithm flow diagram of the proposed method.

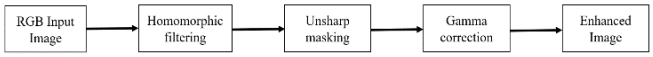


Fig. 3. Algorithm Flow Diagram of Proposed Method.

After getting the RGB input image, the first process applied was homomorphic filtering as it can keep the image details and remove the uneven regions of an image caused by light. The process was followed by unsharp masking to sharpen the image, aimed to retain the details of the image. The final touch on the image would be gamma correction, this technique was applied to further enhance the details of underexposed and overexposed objects while avoiding color distortion problems [33].

The following section will introduce the types of enhancement done on the low-light image.

A. Luminance and Contrast Enhancement

Homomorphic filtering would be used to enhance the luminance and contrast while normalizing the brightness of the image. The idea of this method is to separate illumination and reflectance while applying two different transfer functions to have more control. However, the Fourier transform cannot separate the product of two functions.

$$Fourier[f(x, y)] \neq Fourier[L(x, y)] \times Fourier[R(x, y)] \quad (10)$$

Thus, to apply homomorphic filtering, five steps will be required. First, logarithmic function is applied to the input image, note that the log of the image is expressed as illumination and reflectance of image [34].

$$log[f(x, y)] = log[L(x, y)] + log[R(x, y)] \quad (11)$$

Then, Fast Fourier transform (FFT) is applied on all items where:

$$W_M = e^{-\frac{i2\pi ux}{M}} \quad (12)$$

Applying (12) to Direct Fourier transform (DFT), and its inverse transform, they can be expressed as:

$$Fourier(u) = \sum_{x=0}^{M-1} f(x)\, W_M^{ux},\ u = 0, 1, 2, \dots\dots, M-1 \quad (13)$$

$$InverseFourier(x) = \frac{1}{M}\sum_{u=0}^{M-1} F(u)\, W_M^{-ux},\ x = 0, 1, 2, \dots\dots, M-1 \quad (14)$$

To speed up the Fourier transformation process, FFT algorithm is used instead. Here, the process of analyzing the input frequency of data would be faster as the time complexity of DFT is $O(n^2)$, whereas for FFT the time complexity is $O(n \log n)$. To find the FFT of the image, the equation can be expressed as [35]:

$$F(u, v) = FFT_L(u, v) + FFT_R(u, v) \quad (15)$$

Subsequently, Gaussian filter function, $H(u, v)$, is applied on (15).

$$H(u, v) = (\gamma_H - \gamma_L)\left[1 - e^{-c\frac{\left(\frac{H-u}{2}\right)^2 + \left(\frac{W-v}{2}\right)^2}{D_0^2}}\right] + \gamma_L \quad (16)$$

The default values chosen for the parameters to adjust the Gaussian filter are set to 1.05 for the high-frequency gain, 0.99 for the low-frequency gain, 2 for the constant, and 200 for the cut off frequency.

The constant, c to control the slope steepness, the high-frequency gain, γ_H is set to be greater than 1 and the low-frequency gain, γ_L is set to be lower than 1 to amplify the reflectance of the image while decreasing the illumination and enhancing the contrast of the image. Though there are many suggestions on how these values should be assigned, there are no actual suitable values for these parameters [36]. The values of the parameters can be changed and experimented on to produce a satisfying result. The high pass filter, $H(u, v)$ will then be applied on the Fourier transform, where it will allow high-frequency component while reducing the low-frequency component.

$$S(u, v) = H(u, v)\, FFT_L(u, v) + H(u, v)\, FFT_R(u, v) \quad (17)$$

Followed by Inverse Fast Fourier Transform (IFFT) being applied on the image:

$$s(x, y) = FFT^{-1}\{S(u, v)\} \quad (18)$$

After that, the enhanced illuminance and reflectance of image will be obtained, and to recover the original image, exponential function will be used to reverse the log applied in (11).

$$g(x, y) = exp^{\,s(x, y)} \quad (19)$$

B. Image Details Enhancement

After enhancing the luminance and contrast of the image, the edges of objects in the image would be enhanced so that the details of the image will not lose easily. Thus, unsharp masking would be applied to the image. Box blur method is chosen for the blurring process, Mask is then obtained with the formula of:

$$f_{blurred}(x, y) = \frac{1}{k^2} * \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}_{k^2} \quad (20)$$

Where k here denotes the size of the kernel used for the low-pass filter. Here the value of k is set to 3. Therefore, a 3x3 matrix is used. The reason for using a low kernel size, 3x3 matrix is because the number of pixels that will be blurred will increase as the size of the kernel increases [37] and it would also affect the overall luminosity of the image. Mask is then obtained with the formula of:

$$g_{mask}(x, y) = f(x, y) - f_{blurred}(x, y) \quad (21)$$

After that, result of unsharp masking is obtained using (22), Where $\lambda = 1$ to apply the unsharp mask.

$$g(x, y) = \lambda\,[\, g_{mask}(x, y) + f(x, y)\,] \quad (22)$$

C. Brightness Enhancement

The last step before producing the enhanced image is to perform gamma correction on the enhanced image based on the luminance of the image. In this stage, the gamma value will change according to various luminance to prevent over-enhancement of the already enhanced image. It also controls the overall brightness of the image that is shown on the monitor screen and does not cause serious color distortion.

Here, an algorithm has been proposed to enhance the image where the image will be changed according to the different luminance of the image. The main concept is to separate the luminance of an image into three stages, that is when the luminance is equal to 30, 60, and 120. The reason for separating to three stages is to determine whether the low-light enhancement effect should be low, medium, high, or none. If the low-light image is dark, where the luminance of the image is equal to or lower than 30 then it needs high enhancement, the gamma will be configured to 0.4, else if the low-light image is moderately dark, where the luminance is equal or lower than 60 but higher than 30, the enhancement of the image will be done moderately, where gamma will be set to 0.75. When the luminance of the image is lower than 120, the image will be considered slightly unclear, then a low amount of enhancement will be applied to the low-light image, where the gamma will be set to 0.8. Finally, if the luminance of the image is higher than 120, the image will be considered a clear image, therefore the gamma will be set to 1. The value of gamma is determined empirically after testing various values on the low-light images. Fig. 4 shows the concept of the proposed idea.

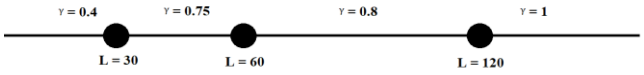


Fig. 4. Concept of Proposed Idea.

Therefore, the pseudocode of the gamma correction applying the proposed concept is:

If the Luminance < = 30,

then set γ = 0.4.

Else if Luminance < = 60,

then set γ = 0.75.

Else if Luminance < 120,

then set γ = 0.8.

Otherwise set γ = 1,

Mathematically, it can be represented as:

$$g(x, y) = \begin{cases} \gamma = 0.4 & if\ L \le 30 \\ \gamma = 0.75 & if\ L \le 60 \\ \gamma = 0.8 & if\ L < 120 \\ \gamma = 1 & otherwise \end{cases} \quad (23)$$

In short, the proposed method was tuned to enhance the luminance, contrast, image details and brightness of the image.

V. Experimental Results

TID2013 dataset [38] has been chosen to evaluate the potential and effects of the proposed method. MATLAB version R2021b running on a laptop with Intel Core i5-8250u CPU operating at 1.60 GHz with physical memory of 8.00 GB was chosen to obtain the results. A comparison between the proposed method with the LIME [39] method and DYNENH [40] method has been made. A visual comparison between the results of "Wall", "Caps", "Portrait", "Barn", "Forest", "Airplane", "Lighthouse", and "Flower" is listed in Fig. 5.

A. Visual Analysis

Visual comparison made in Fig. 5 showed that the images that were enhanced using the LIME method caused an over-enhancement problem and caused color distortion and loss of image details in the enhanced image. The color distortion can be observed in the overly enhanced image of the lighthouse where in the ground truth, the roof of the house beside is dark brown whereas, in the enhanced image, the color of the roof has turned bright and reddish-brown color. Besides, the loss of image details occurred in the LIME method as there are some figures in the lighthouse image which was unable to be captured clearly. Another apparent image that showed an over-enhancement problem with the LIME method is the airplane. The color of the letters 'SIX-SHOOTER' has become a sky-blue color instead of the deep blue color in its ground truth. Forest image has also shown an obvious color distortion after being enhanced with the LIME method, where the enhanced image shows bright yellowish-green bushes and trees compared to the original dull green color. On the other hand, the DYNENH method gives rise to the problem of the loss of details in the image. This can be seen in the airplane image, where the low-light image is not properly enhanced, making the image turn dark and blurry. After getting the input of low-light image, homomorphic filtering will be applied to the low-light image, enhancing the luminance of the whole image, followed by unsharp masking that enhances the contrast of the image and sharpens the edges of the image so that the details of the image will be seen clearer, finally, the gamma correction will be done on the image according to the image's luminance, gamma if then arranged according to the luminance of the image, as proposed in our method.

Fig. 5 shows the proposed method had successfully minimize the problem of color distortion and loss of image details while enhancing low-light images. In addition, the color and brightness of the enhanced image are similar to the ground truth.

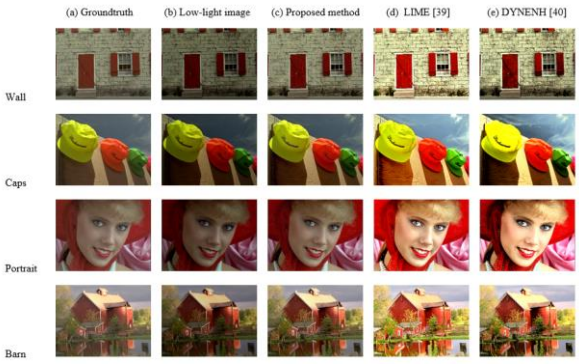


Fig. 5. Visual Comparison between Enhanced Low-Light Images.

B. Image Details Preservation

An objective evaluation has been chosen to compare the differences between the three low-light image enhancement methods, which include Features Similarity Index (FSIM), Features Similarity Index color (FSIMc) [41], Mean Square Error (MSE), and Structural Similarity Index (SSIM) [42]. FSIM and FSIMc have proven to be effective and consistent in evaluating image quality by measuring the image chromatic features, and higher FSIM and FSIMc indicate that an enhanced image is more similar to the ground truth. MSE is a full-reference quality metric that finds the average of squared intensity differences of distorted and reference image pixels. The smaller MSE represents small errors, and the enhanced image is similar to the ground truth. Besides that, SSIM is used to compare normalized pixel intensities' patterns, therefore a higher SSIM value represents the higher similarity of structural features between the ground truth and the enhanced image [42],[43],[44].

Table I shows the evaluation result for the preservation of image details. The proposed method has proven to be best as compared to the LIME and DYNENH methods. The FSIMc, FSIM, and SSIM are the highest among all the compared methods which indicate that the proposed method can preserve the details of the image better when compared to the other two. The proposed method has adjusted the luminance and contrast of the image using the frequency domain method and applied the unsharp masking method to enhance the edges of the object in the image, therefore producing enhanced images that are similar to their originals. Besides that, the proposed method also has the lowest MSE values when compared with the other two methods.

C. Naturalness Preservation and Visual Information

The naturalness preservation and visual information of the enhanced image were also evaluated. Lightness order error (LOE) has been used to assess the naturalness preservation of the enhanced image [45]. The lower LOE represents better naturalness preservation as LOE shows the value of lightness distortion in the image. Visual information fidelity (VIF) can measure the accuracy which relates to the quality of visual information that is perceived by the human visual system [46],[47]. It helps in identifying the distortion of visual information in the enhanced image, therefore a higher VIF means a better image quality. Table II shows the value of LOE and VIF of the images for each method.

TABLE I. EVALUATION ON IMAGE DETAILS PRESERVATION

Image	Evaluation	Proposed Method	LIME [39]	DYNENH [40]
Wall	FSIMc	**0.9504**	0.7750	0.8132
	FSIM	**0.9522**	0.7851	0.8176
	MSE	**266.0**	5131.9	2049.8
	SSIM	**0.8815**	0.6964	0.7106
Caps	FSIMc	**0.9714**	0.8334	0.7995
	FSIM	**0.9731**	0.8443	0.8088
	MSE	**212.6**	2595.4	3105.4
	SSIM	**0.9163**	0.7873	0.6715
Portrait	FSIMc	**0.9688**	0.8509	0.8062
	FSIM	**0.9701**	0.8617	0.8155
	MSE	**214.6**	2223.9	3484.2
	SSIM	**0.8975**	0.8006	0.7023
Barn	FSIMc	**0.9599**	0.7904	0.8374
	FSIM	**0.9611**	0.7998	0.8428
	MSE	**183.0**	2590.3	1339.1
	SSIM	**0.9042**	0.7432	0.7560
Forest	FSIMc	**0.9562**	0.7284	0.7910
	FSIM	**0.9575**	0.7382	0.7983
	MSE	**296.0**	4576.7	2531.5
	SSIM	**0.8661**	0.6227	0.7123
Airplane	FSIMc	**0.9652**	0.8424	0.8372
	FSIM	**0.9659**	0.8477	0.8383
	MSE	**339.4**	2009.2	4805.8
	SSIM	**0.8645**	0.8119	0.5295
Lighthouse	FSIMc	**0.9560**	0.7985	0.8222
	FSIM	**0.9570**	0.8054	0.8237
	MSE	**195.8**	3148.7	1798.2
	SSIM	**0.9125**	0.7771	0.6343
Flower	FSIMc	**0.9604**	0.7467	0.8064
	FSIM	**0.9617**	0.7547	0.8119
	MSE	**190.1**	4984.0	2475.6
	SSIM	**0.9238**	0.7018	0.7106

TABLE II. EVALUATION ON LUMINANCE AND VISUAL INFORMATION DISTORTION

Image	Evaluation	Proposed Method	LIME [39]	DYNENH [40]
Wall	LOE	**381.1**	1524.0	958.6
	VIF	**0.6030**	0.5089	0.4858
Caps	LOE	**120.8**	504.9	462.4
	VIF	**0.6807**	0.4613	0.3683
Portrait	LOE	**258.2**	955.9	800.4
	VIF	**0.6837**	0.5171	0.4367
Barn	LOE	**282.0**	1381.1	423.3
	VIF	**0.6651**	0.4822	0.4387
Forest	LOE	**317.4**	1653.0	475.9
	VIF	**0.5954**	0.3521	0.4124
Airplane	LOE	**204.7**	2075.1	702.5
	VIF	**0.6166**	0.4010	0.3722
Lighthouse	LOE	**435.4**	1016.7	603.9
	VIF	**0.6066**	0.4117	0.3096
Flower	LOE	**217.1**	1488.7	612.0
	VIF	**0.6613**	0.4554	0.3920

According to the results, the proposed method has the lowest LOE and highest VIF among all the other methods, proving that it can enhance low-light images and produce a better image with the lowest value of lightness distortion and highest accuracy compared to the other two methods.

VI. DISCUSSION

The luminance and contrast of low-light image was enhanced during the homomorphic filtering, which had the advantage of maintaining details of image while also removing the uneven regions of images that were caused by light. To tackle the issue of loss of image details, unsharp masking would be used to sharpen the image and produce clearer edge for the object. This step also prevents the loss of the object edges in the image on the following step, that is gamma correction. Gamma correction would show the details of objects that are underexposed or overexposed while avoiding color distortion problem. The results shown in Section V had proved how the proposed method excelled than the other two existing methods. The proposed method had successfully preserved the details of image, achieving the highest FSIMc, FSIM values compared to the other two existing methods, that is, above 0.9, showing its effectiveness in recovering the chromatic features of the image. Besides that, it also achieved the SSIM values above 0.8, showing the similarity of the pixels' pattern between the enhanced image and the ground truth. Additionally, the proposed method managed to have lowest MSE among the three, reaching the lowest value of 183.0 in the "Barn" image. Homomorphic filtering and unsharp masking have contributed a lot in maintaining the details of the images. Aside from preserving the details of the image, the proposed method also showed its capability in enhancing the low-light image while preserving the naturalness of image and visual information. It managed to achieve lowest LOE and highest VIF scores. Gamma correction played a big role in showing the details of underexposed and overexposed objects while avoiding the color distortion problems, which led to a more natural looking image. Therefore, all three processes involved in the proposed method are important in maintaining the chromatic features of image, its structure, as well as enhancing the image and producing a natural looking image. The proposed method in this research has successfully enhanced a low-light image and produced an enhanced image with minimal color distortion and with clear details. The results showed that the proposed method has better visual quality compared to other methods. The proposed method also achieved the lowest MSE and LOE scores, and the highest FSIMc, FSIM, SSIM, and VIF scores. This proves that the proposed method has the lowest color distortion and is the best in preserving the details of the image, thus outperforming the other two low-light image enhancement methods.

VII. CONCLUSION

This research has successfully created a method in the gamma correction stage that enhances the low-light image using the concept of separating the luminance of the image into three stages and enhancing them with specified gamma according to the stage it belongs to. Compared to other low-light enhancement methods, the proposed method performs better in minimizing the color distortion errors, retaining the image details, and producing a more natural enhanced image. Additionally, the proposed method can preserve features of the image better compared to the other two low-light image

enhancement methods. Although the proposed method can enhance the image with minimal color distortion and is able to produce an image with clear details, it is not tuned to remove the haze of a low-light image, and this might cause difficulties when the low-light image is taken in a hazy environment. Besides that, since the parameters of the values for the high pass filter are determined empirically, they have to be adjusted manually rather than using the default values to produce a satisfying result. As the image enhancement becomes more important in digital images, there will be a lot of improvements made to the proposed method before it can be applied in daily lives. The future work will be focused on improving the method so that it can dehaze and reduce the noise in low-light images; these functionalities are important as low-light images might be captured in a hazy situation, for an instant a hazy day, and a noisy low-light image should be considered. Another thing to work on is to upgrade the method so that it can be applied in real-time applications. The method had the drawback of being unable to utilize the best high pass filter for the low-light image since the values of parameters for the high pass filter had to be adjusted manually for some images. In the future, an algorithm should be implemented for the high pass filter so that it can automatically find the best values for the parameters of the high pass filter.

ACKNOWLEDGMENT

Authors are thankful to Universiti Sultan Zainal Abidin, Terengganu Darul Iman, Malaysia for the financial funding and Fakulti Informatik dan Komputeran for supporting us in accomplishing this study.

REFERENCES

[1] W. Wang, X. Wu, X. Yuan, and Z. Gao, "An Experiment-Based Review of Low-Light Image Enhancement Methods," *IEEE Access*, vol. 8, pp. 87884–87917, 2020, doi: 10.1109/access.2020.2992749.

[2] L. Zhuang and Y. Guan, "Image Enhancement by Deep Learning Network Based on derived image and Retinex," 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Oct. 2019, doi: 10.1109/imcec46724.2019.8983874.

[3] Z. Gu, F. Li, F. Fang, and G. Zhang, "A Novel Retinex-Based Fractional-Order Variational Model for Images With Severely Low-light," IEEE Transactions on Image Processing, vol. 29, pp. 3239–3253, 2020, doi: 10.1109/tip.2019.2958144.

[4] D. Jobson, Z. Rahman, and G. Woodell, "Properties and performance of a center/surround Retinex," IEEE Transactions on Image Processing, vol. 6, no. 3, pp. 451-462, 1997, doi: 10.1109/83.557356.

[5] D. Jobson, Z. Rahman, and G. Woodell, "Multi-scale Retinex for color image enhancement," IEEE International Conference on Image Processing, pp. 1003-1006, 1996, doi: 10.1109/ICIP.1996.560995.

[6] D. Jobson, Z. Rahman, and G. Woodell, "A multiscale Retinex for bridging the gap between color images and the human observation of scenes," IEEE Transactions on Image Processin, vol. 6, no. 7, pp. 965-976, 2002, doi: 10.1109/83.597272.

[7] Z. Rahman, D. Jobson, and G. Woodell, "Retinex processing for automatic image enhancement," Journal of Electronic Imaging, vol. 13, no. 1, pp. 100-110, 2004. doi: 10.1117/1.1636183.

[8] S. Liu, W. Long, L. He, Y. Li, and W. Ding, "Retinex-Based Fast Algorithm for Low-Light Image Enhancement," Entropy, vol. 23, no. 6, p. 746, Jun. 2021, doi: 10.3390/e23060746.

[9] S. Wang, D. Gao, Y. Wang, and S. Wang, "An Improved Retinex low-illumination image enhancement algorithm," Proceedings of APSIPA Annual Summit and Conference 2019, pp. 1134–1139.

[10] Y. Shi, X. Wu, and M. Zhu, "Low-light Image Enhancement Algorithm Based on Retinex and Generative Adversarial Network," Jun. 2019, doi: 10.48550/arXiv.1906.06027.

[11] L. H. Pham, D. N.-N. Tran, and J. W. Jeon, "Low-Light Image Enhancement for Autonomous Driving Systems using DriveRetinex-Net," 2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), Nov. 2020, doi: 10.1109/icce-asia49877.2020.9277442.

[12] P. K. Mishro, S. Agrawal, R. Panda, and A. Abraham, "A novel brightness preserving joint histogram equalization technique for contrast enhancement of brain MR images," Biocybernetics and Biomedical Engineering, vol. 41, no. 2, pp. 540–553, Apr. 2021, doi: 10.1016/j.bbe.2021.04.003.

[13] U. K. Acharya and S. Kumar, "Genetic algorithm based adaptive histogram equalization (GAAHE) technique for medical image enhancement," Optik, vol. 230, p. 166273, Mar. 2021, doi: 10.1016/j.ijleo.2021.166273.

[14] P. Kaur, B. S. Khehra, and A. P. S. Pharwaha, "Color Image Enhancement based on Gamma Encoding and Histogram Equalization," Materials Today: Proceedings, vol. 46, pp. 4025–4030, 2021, doi: 10.1016/j.matpr.2021.02.543.

[15] S. D. Chen and A. R. Ramli, "Minimum mean brightness error bi-histogram equalization in contrast enhancement," IEEE Transactions on Consumer Electronics, vol. 49, no. 4, pp. 1310–1319, Nov. 2003, doi: 10.1109/tce.2003.1261234.

[16] A. M. Reza, "Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement," The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, vol. 38, no. 1, pp. 35–44, Aug. 2004, doi: 10.1023/b:vlsi.0000028532.53893.82.

[17] L. Zhuang and Y. Guan, "Image Enhancement via Subimage Histogram Equalization Based on Mean and Variance," Computational Intelligence and Neuroscience, vol. 2017, pp. 1–12, 2017, doi: 10.1155/2017/6029892.

[18] C. K. Sirajuddeen, S. Kansal, and R. K. Tripathi, "Adaptive histogram equalization based on modified probability density function and expected value of image intensity," Signal, Image and Video Processing, vol. 14, no. 1, pp. 9–17, Jun. 2019, doi: 10.1007/s11760-019-01516-2.

[19] Y. Qi, Z. Yang, J. Lian, Y. Guo, W. Sun, J. Liu, R. Wang and Y. Ma, "A new heterogeneous neural network model and its application in image enhancement," Neurocomputing, vol. 440, pp. 336–350, Jun. 2021, doi: 10.1016/j.neucom.2021.01.133.

[20] C. Vimala and P. A. Priya, "Artificial neural network-based wavelet transform technique for image quality enhancement," Computers & Electrical Engineering, vol. 76, pp. 258–267, Jun. 2019, doi: 10.1016/j.compeleceng.2019.04.005.

[21] P. Gómez, M. Semmler, A. Schützenberger, C. Bohr, and M. Döllinger, "Low-light image enhancement of high-speed endoscopic videos using a convolutional neural network," Medical & Biological Engineering & Computing, vol. 57, no. 7, pp. 1451–1463, Mar. 2019, doi: 10.1007/s11517-019-01965-4.

[22] E. Ha, H. Lim, S. Yu, and J. Paik, "Low-light Image Enhancement Using Dual Convolutional Neural Networks for Vehicular Imaging Systems," 2020 IEEE International Conference on Consumer Electronics (ICCE), Jan. 2020, doi: 10.1109/icce46568.2020.9043035.

[23] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hoi, S. Kwong and R. Cong, "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement," Mar. 2020, doi: https://doi.org/10.48550/arXiv.2001.06826.

[24] Y. Guo, X. Ke, J. Ma, and J. Zhang, "A Pipeline Neural Network for Low-Light Image Enhancement," IEEE Access, vol. 7, pp. 13737–13744, 2019, doi: 10.1109/access.2019.2891957.

[25] A. Kumar, R. K. Jha, and N. K. Nishchal, "An improved Gamma correction model for image dehazing in a multi-exposure fusion framework," Journal of Visual Communication and Image Representation, vol. 78, p. 103122, Jul. 2021, doi: 10.1016/j.jvcir.2021.103122.

[26] S. Zaheeruddin and K. Suganthi, "Image Contrast Enhancement by Homomorphic Filtering based Parametric Fuzzy Transform," Procedia

Computer Science, vol. 165, pp. 166–172, 2019, doi: 10.1016/j.procs.2020.01.095.

[27] S. Adhikari and S. P. Panday, “Image Enhancement Using Successive Mean Quantization Transform and Homomorphic Filtering,” 2019 Artificial Intelligence for Transforming Business and Society (AITB), Nov. 2019, doi: 10.1109/aitb48515.2019.8947437.

[28] M. Veluchamy and B. Subramani, “Image contrast and color enhancement using adaptive gamma correction and histogram equalization,” Optik, vol. 183, pp. 329–337, Apr. 2019, doi: 10.1016/j.ijleo.2019.02.054.

[29] P. Li, J. Liang, and M. Zhang, “A degradation model for simultaneous brightness and sharpness enhancement of low-light image,” Signal Processing, vol. 189, p. 108298, Dec. 2021, doi: 10.1016/j.sigpro.2021.108298.

[30] R. C. Gonzalez and R. E. Woods, Digital image processing. New York, Ny: Pearson, 2018.

[31] E. Baidoo, “Implementation of Gray Level Image Transformation Techniques,” International Journal of Modern Education and Computer Science, vol. 10, no. 5, pp. 44–53, May 2018, doi: 10.5815/ijmecs.2018.05.06.

[32] J. C. González and Ò. Daniel Carmona Salazar, “Image Enhancement with Matlab Algorithm,” 2015, Accessed: Jan. 18, 2022. [Online]. Available: http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A817007&dswid=7828.

[33] W. Wang, X. Yuan, Z. Chen, X. Wu, and Z. Gao, “Weak-Light Image Enhancement Method Based on Adaptive Local Gamma Transform and Color Compensation,” Journal of Sensors, vol. 2021, pp. 1–18, Jun. 2021, doi: 10.1155/2021/5563698.

[34] M. N. Haque & M. S. Uddin, “Accelerating Fast Fourier Transformation for Image Processing using Graphics Processing Unit.” Journal of Emerging Trends in Computing and Information Sciences. 2. 367-375. Jan. 2011.

[35] Z. Shi, Y. Chen, E. Gavves, P. Mettes, and C. G. M. Snoek, “Unsharp Mask Guided Filtering,” IEEE Transactions on Image Processing, vol. 30, pp. 7472–7485, 2021, doi: 10.1109/tip.2021.3106812.

[36] A. M. Saleh, Sami “Mathematical Equations for Homomorphic Filtering in Frequency Domain: A Literature Survey,” 2012 International Conference of Information and Knowledge Management (ICIKM), IACSIT Press, Singapore, vol. 45, pp. 74–77, 2012.

[37] E. Pulfer “Different Approaches to Blurring Digital Images and Their Effect on Facial Detection” Computer Science and Computer Engineering Undergraduate Honors Theses, May 2019, Retrieved from https://scholarworks.uark.edu/csceuht/66.

[38] N. Ponomarenko et al., “Image database TID2013: Peculiarities, results and perspectives,” Signal Processing: Image Communication, vol. 30, pp. 57–77, Jan. 2015, doi: 10.1016/j.image.2014.10.009.

[39] X. Guo, Y. Li, and H. Ling, “LIME: Low-Light Image Enhancement via Illumination Map Estimation,” IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 982–993, Feb. 2017, doi: 10.1109/tip.2016.2639450.

[40] G. Li, M. N. A. Rana, J. Sun, Y. Song, and J. Qu, “Real-time image enhancement with efficient dynamic programming,” Multimedia Tools and Applications, vol. 79, no. 41–42, pp. 30883–30903, Aug. 2020, doi: 10.1007/s11042-020-09586-y.

[41] Lin Zhang, Lei Zhang, Xuanqin Mou, and D. Zhang, “FSIM: A Feature Similarity Index for Image Quality Assessment,” IEEE Transactions on Image Processing, vol. 20, no. 8, pp. 2378–2386, Aug. 2011, doi: 10.1109/tip.2011.2109730.

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/tip.2003.819861.

[43] A. Z. Abd Aziz and H. Wei, "Polarization Imaging for Face Spoofing Detection: Identification of Black Ethnical Group," 2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA), 2018, pp. 1-6, doi: 10.1109/ICASSDA.2018.8477608.

[44] S. Wang, J. Zheng, H.-M. Hu, and B. Li, “Naturalness Preserved Enhancement Algorithm for Non-Uniform Illumination Images,” IEEE Transactions on Image Processing, vol. 22, no. 9, pp. 3538–3548, Sep. 2013, doi: 10.1109/tip.2013.2261309.

[45] H. R. Sheikh, A. C. Bovik, “A visual information fidelity approach to video quality assessment,” The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics, vol. 7, 2015.

[46] M. Moniruzzaman, M. A. K. Hawlader, M. F. Hossain and M. A. Rashid, "SVD and chaotic system based watermarking approach for recovering crime scene image," 8th International Conference on Electrical and Computer Engineering, 2014, pp. 132-135, doi: 10.1109/ICECE.2014.7026987.

[47] Z. Mohamad, L. Y. Thong, A. H. Zakaria and W. S. W. Awang, "Image based authentication using zero-knowledge protocol," 2018 4th International Conference on Computer and Technology Applications (ICCTA), 2018, pp. 202-210, doi: 10.1109/CATA.2018.8398683.

An Ensemble of Arabic Transformer-based Models for Arabic Sentiment Analysis

Ikram El Karfi, Sanaa El Fkihi
ENSIAS, Mohammed V University, Rabat, Morocco

***Abstract*—In recent years, sentiment analysis has gained momentum as a research area. This task aims at identifying the opinion that is expressed in a subjective statement. An opinion is a subjective expression describing personal thoughts and feelings. These thoughts and feelings can be assigned with a certain sentiment. The most studied sentiments are positive, negative, and neutral. Since the introduction of attention mechanism in machine learning, sentiment analysis techniques have evolved from recurrent neural networks to transformer models. Transformer-based models are encoder-decoder systems with attention. Attention mechanism has permitted models to consider only relevant parts of a given sequence. Making use of this feature in encoder-decoder architecture has impacted the performance of transformer models in several natural language processing tasks, including sentiment analysis. A significant number of Arabic transformer-based models have been pre-trained recently to perform Arabic sentiment analysis tasks. Most of these models are implemented based on Bidirectional Encoder Representations from Transformers (BERT) such as AraBERT, CAMeLBERT, Arabic ALBERT and GigaBERT. Recent studies have confirmed the effectiveness of this type of models in Arabic sentiment analysis. Thus, in this work, two transformer-based models, namely AraBERT and CAMeLBERT have been experimented. Furthermore, an ensemble model has been implemented to achieve more reasonable performance.**

Keywords*—*Transformers; BERT; ensemble learning; Arabic sentiment analysis

I. Introduction

Given the tremendous amounts of Arabic digital content that has been produced in the last couple of years, an increasing number of research works have been devoted to the automatic processing of this language. In this regard, different techniques have been used to classify a specific text. Many studies have addressed this task by making use of basic machine learning models such as Naïve Bayes (NB) and Support Vector Machine (SVM). The authors in [1] addressed Arabic text classification using SVM and NB combined with the N-gram feature. The best accuracy of SVM was achieved without the N-gram, as for NB the best accuracy was achieved when the N-gram feature was considered. Whereas the authors in [2] introduced their Arabic Jordanian twitter corpus, then evaluated N-grams and stemming techniques together with TF-IDF or TF weighting schemes. Experiments have been carried out by making use of SVM and NB. Results showed that training SVM model on top of stems and bigrams using TF-IDF could give better performance compared to NB model. In a similar work [3], the authors performed sentiment analysis on Arabizi text which is Arabic text written in Latin alphabets. For experimentation purposes, the authors used NB and SVM classifiers. Besides, they evaluated the filtering step, which consists of removing stop words and mapping emojis to their corresponding Arabic words. Results indicated that SVM model outperformed NB model. However, filtering step did not greatly improve the accuracy.

Recently deep learning models such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) have proven to be efficient for analyzing Arabic content. Many researchers have relied on deep learning models to tackle Arabic sentiment analysis task. In [4] performed a binary sentiment classification in Arabic. Initially, they applied preprocessing to clean input texts. Next, a word embedding layer has been used to represent texts as numerical vectors to be fed to the LSTM layer. Finally, a SoftMax layer followed to predict the polarity of the text. The experiments showed quite good results with an accuracy ranging from 80% to 82%. In another work [5] the authors made an empirical comparison between deep learning models (LSTM, CNN) and other machine learning models for both binary and multiclass classification using different datasets. The paper showed that deep learning models are effective for larger datasets. In contrast, basic machine learning algorithms perform well on smaller datasets.

More recently, with the increasing popularity of transformer models, sentiment analysis task has been significantly improved in terms of performance. Transformer models have replaced deep learning models and achieved state-of-the-art results on many automatic language processing tasks such as sentiment analysis [6], named entity recognition [7], question answering [8], and many other tasks.

The ineffectiveness of the existing methods performing sentiment analysis in Arabic is the main motivation for proposing a transformer-based ensemble method. In the last few years transformer language models alone led to significant improvements in sentiment analysis. Hence, making use of the advantages of this type of models to investigate more reliable approaches is indeed necessary to tackle sentiment analysis in Arabic being a morphologically rich language. In this paper, we propose an ensemble model that combines the strengths of two transformer language models.

Many different disciplines have made significant use of ensemble approaches to address text classification. However, there is relatively few studies on the use of ensemble methods in Arabic sentiment analysis. The primary goal in this paper is to propose an ensemble model that combines two base transformer models namely AraBERT and CAMeLBERT into a single model. To the best of our knowledge this is the first

study that investigates the ensemble of transformer-based models in Arabic sentiment analysis. Experimental results demonstrate that the proposed ensemble method outperforms stand-alone classifiers and majority voting ensemble model.

The rest of this paper is structured as follows. Related works will be introduced in Section II. The overall proposed methodology will be discussed in Section III. Experiments and results are given in Section IV. Then conclusions are drawn in Section V.

II. Related Work

Given the effectiveness of transformer-based models, there have been various transformer models used in Arabic sentiment analysis. The widely utilized models are Multilingual BERT, AraBERT, and MARBERT [9]. The author in [10] addressed sentiment analysis in Modern Standard Arabic (MSA) and other Arabic dialects such as Levantine, Egyptian, and Gulf. The author opted for three-way classification according to three scales (positive, neutral, and negative) and using different algorithms, namely: Naive Bayes classifiers (NB), Support Vector Machine (SVM), Random Forest Classifier, and BERT model (Bidirectional Encoder Representations from Transformers). The best results are obtained with BERT model reaching an accuracy score of more than 83%. The author in [11] addressed sarcasm and sentiment detection using two variants of transformer-based models, namely AraELECTRA and AraBERT. Evaluation results showed that AraBERT performs the best in terms of accuracy for both sarcasm and sentiment detection. In a similar work, [12] addressed the same tasks: sarcasm detection and sentiment analysis. The authors have examined six BERT-based models including: MARBERT [13], QARiB [14], AraBERTv02 [15], GigaBERTv3 [16], Arabic BERT [17], and mBERT [18]. MARBERT achieved promising results for both tasks.

Several studies in the literature investigated ensemble methods in Arabic sentiments analysis. The authors in [19] investigated different deep learning models to improve Arabic sentiment analysis accuracy. The authors proposed an ensemble model combining a Convolutional Neural Network (CNN) model and a Long Short-Term Memory (LSTM) model. To evaluate their model, they used the ASTD dataset [20] which consists of 10000 tweets. In this work, they focused only on opinion classification, hence the objective class tweets were removed. To construct their ensemble model, they experimented different CNN models and LSTM models with different hyper-parameters. The best CNN model is obtained by configuring the parameter fully connected layer size to 100. As for LSTM, the best model is obtained by using a dropout rate of 0.2, based on the best CNN model and the best LSTM model they built an ensemble model where the final predicted class is obtained using soft voting. Results show that the ensemble model achieved better results in terms of accuracy and F1-score compared to LSTM model and CNN model. In another study [21] , the authors introduced their approach to address three SemEval related sentiment analysis subtasks in Arabic. First Subtask (A) addresses Message Polarity Classification, then Subtask (B) addresses Topic-Based Message Polarity Classification, finally Subtask (D) which addresses Tweet quantification. The authors proposed two systems, the first is developed using their previous proposed sentiment analyzer [22] based on a scored lexicon. The second system is an ensemble of three different classifiers namely Convolutional Neural Network using Word2vec, Multilayer Perceptron and Logistic Regression. Using voting between the three classifiers the authors determined the final outcome. Evaluation results showed that their systems outperformed all the other systems by achieving an accuracy of 0.58 and 0.77 on Subtask A and Subtask B respectively, as for Subtask D their system outperformed the other systems as well by achieving 0.127 in terms of KLD score.

It seems clear that none of the existing ensemble methods has addressed Arabic sentiment analysis by making use of transformer language models. Accordingly, this study will be focusing on investigating the use of transformer language models in Arabic sentiment classification as well as proposing an ensemble technique based on transformer models to enhance classification accuracy.

III. Proposed Methodology

This section presents the methodology proposed in this paper. We will be discussing the background of transformer-based models and the models adopted in this work. Then, describe the proposed ensemble model architecture.

A. Background

A variety of neural network architectures have been proposed and used for text classification tasks, including sentiment classification. Among these numerous architectures, the best adapted architecture to sequential data is recurrent neural networks (RNNs). They have demonstrated to be effective on data where elements order is important. For example, in a sentence, the order in which words occur has a significant impact on the meaning of that sentence. Since its introduction, RNNs have been the state-of-the-art for capturing and processing dependencies in sequences. Nevertheless, it also has its drawbacks, it has been proved that RNNs cannot process large sequences of text such as long paragraphs. Moreover, in practice, data is processed sequentially, which makes it difficult to perform parallel computing using RNNs. Recently, a new architecture called Transformers has been introduced. Similar to RNNs, transformers use attention mechanism and inherit the encoder-decoder architecture of the sequence-to-sequence models to deal with sequential data. However, its architecture does not involve recurrent networks in order to speed up the training process. Transformers were firstly introduced by [23], and they were initially designed to perform translation. As illustrated below in Fig. 1, a transformer consists of two blocks, on the left, the encoder stack, and on the right, the decoder stack. The encoder stack is made up of a multi-head self-attention layer and a fully connected feed forward network. In addition to these two layers, the decoder stack has one more layer called the masked multi-head attention layer. As transformer architecture does not rely on recurrence, word position is not provided. To preserve this information positional encoding technique has been introduced. In addition to the input embedding vector, a positional vector with the same dimension as the input embedding vector is added to capture the context of a word in a sentence.

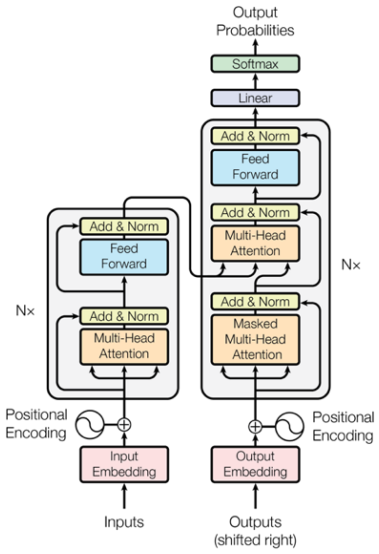


Fig. 1. Transformer Architecture [23].

Transformer-based models include three types of models: encoder-only, decoder-only, and encoder-decoder, following a brief introduction to each type of transformers, its architecture and its applications.

1) Encoder-only models: In this type of transformers only the encoder part is needed. A vector representing the input sequence is fed to the first encoder block that consists of a bi-directional self-attention layer and a feed-forward layer, the output of this block is passed to the following encoder block, which itself is composed of two layers. Each encoder block tries to enrich the embedding vector with contextual information. The final encoder block outputs the last contextual encoding. This type of transformer is suitable for tasks such as text classification or named entity recognition. The most popular examples of this type of models are: BERT [18], ELECTRA [24], and RoBERTA [25].

2) Decoder-only models: In decoder models only decoder stack is used. It consists of N identical decoder blocks; a single decoder block is composed of three layers. The first layer is a masked multi-head attention layer, in which future information is masked and only previous positions in the input sequence have attention. Similarly, to the encoder block, the next layers are multi-head self-attention layers and a fully connected feed-forward network. Decoder-only based models are also called autoregressive models and are more suited for tasks such as text generation. The most widely used models trained with decoder-only architectures are GPT (Generative Pre-trained Transformer) [26].

3) Encoder-decoder models: Also called sequence-to-sequence models, this type of models is implemented using both blocks: encoder block and decoder block. In the encoder block, the whole sequence is considered, whereas in the decoder block for a given word, only the words that precede are considered. Encoder-decoder models are best suited for tasks that involve the input of a sequence of items (words, letters, etc.) and then outputs another sequence. This architecture can be applied in the case of machine translation or question answering, where a sequence of words is treated sequentially and the result is also a sequence of words. Recently, many encoder-decoder based models have been introduced.

B. Transformer Language Models for Arabic

Transformers were initially introduced as a novel architecture for translation. Ever since, they have been mostly used for natural language processing. In sentiment analysis task, pretrained transformer language architectures have significantly improved the performance of models. Each model has its own size and trained on different type of datasets. Table I summarizes the most applied models in Arabic text classification.

In this paper, we have selected some of the most effective Arabic transformation models in sentiment analysis in Arabic. Each of these models is based on different architectures and trained using different Arabic variants. Hereafter, we discuss each of the selected models and their architecture.

1) AraBERT [27] pretrained BERT model using a pretrained dataset of 70 million sentences, collected from Wikipedia dumps, Arabic news websites and two large corpora: Abulkhair Arabic Corpus [31] and OSIAN [32]. AraBERT comes in two versions AraBERTv0.1 and AraBERTv1. In this study AraBERTv0.2 is used for experiments.

TABLE I. SUMMARY OF ARABIC PRETRAINED MODELS

Model name	Ref	Size	Source	Data type	Parameters
Multilingual BERT	[18]	-	Wikipedia	MSA	110M
AraBERT	[27]	24GB	Wikipedia+ Abulkhair Corpus+ OSIAN+ news websites	MSA	136M(base) 371M(large)
ArabicBERT	[17]	95GB	Wikipedia+ OSCAR+ other sources	MSA/ Dialect	110M(base) 340M(large)
CAMeLBERT	[28]	167B	Gigaword+ Abulkhair Corpus+ OSIAN+ Wikipedia+ OSCAR+ dialectal corpora+ OpenITI corpus	MSA / Dialect/ Classical	17.3B
MARBERT	[13]	128GB	Twitter API	MSA/ Dialect	160M
Arabic ALBERT	[29]	-	OSCAR+ Wikipedia	MSA	12M(base) 18M(large) 60M(xlarge)
GigaBERT	[16]	-	Gigaword+ Wikipedia+ OSCAR	MSA	125M
XLM-RoBERTa	[30]	2.5TB	CommonCrawl	MSA	270M(base) 550M(large)

2) CAMeLBERT [28] implemented their Arabic pre-trained language model on top of three variants of Arabic: Modern Standard Arabic (MSA), dialectal Arabic, and classical Arabic. The authors evaluated the proposed model by making use of 12 datasets to address five tasks: Named Entity Recognition, POS tagging, sentiment analysis, dialect identification, and poetry classification.

C. Ensemble Models

Ensemble learning is a technique that combines multiple machine learning models to improve the performance of the learning model and achieve a higher accuracy score than would be achieved by any single model in the ensemble. In this study, two ensemble techniques have been evaluated. The first technique is the majority voting. It is the most commonly used technique for ensemble learning. The second technique is based on calculating the sum of raw outputs of each model in the ensemble.

1) Majority voting: In majority voting, the final output of the ensemble model is determined by counting for each class the number of votes of multiple models. The class with the majority of votes is predicted.

2) The proposed method SUM: As illustrated in Fig. 2, that represents the proposed ensemble model. Firstly, a raw text is fed to the model as input, then transformed into vector representation so that it can be processed with encoder-decoder approach. Then the decoder-block of each model outputs probabilities for each class. Finally, the output is obtained by calculating the weighted sum of the probabilities from the same class, then for each class, the argmax operation is applied to find the class with higher probability value.

For a given text, let PAr_{Neg} and PAr_{Pos} denote the probabilities predicted with AraBERT model for the class Negative and the class Positive respectively. Whereas, $PCaM_{Neg}$ and $PCaM_{Pos}$ denote the probabilities predicted with CAMeLBERT model for the class Negative and the class Positive respectively. For each class, the final probability is obtained by calculating the weighted sum of both probabilities, namely the probability given with AraBERT model and CAMeLBERT model. Weights values are not selected randomly, the main reason of selecting weight values 0, 7 and 0, 3 for CAMeLBERT and AraBERT respectively, is that CAMeLBERT model tend to perform well on the majority of the datasets (see Table II). Thus, we considered 70% of the probability generated by CAMeLBERT model and 30% of the probability generated by AraBERT model. The final probabilities are calculated using the following equations:

$$PF_{Neg} = (0.3 \times PAr_{Neg}) + (0.7 \times PCaM_{Neg})$$

$$PF_{Pos} = (0.3 \times PAr_{Pos}) + (0.7 \times PCaM_{Pos})$$

Next, the final output corresponds to the index with higher probability value.

$$\text{Final Output} = \text{argmax}([PF_{Neg}, PF_{Pos}])$$

Therefore, if the output index is 0, the model will assign to the input text the class Negative, and if the output index is 1 the model will assign Positive.

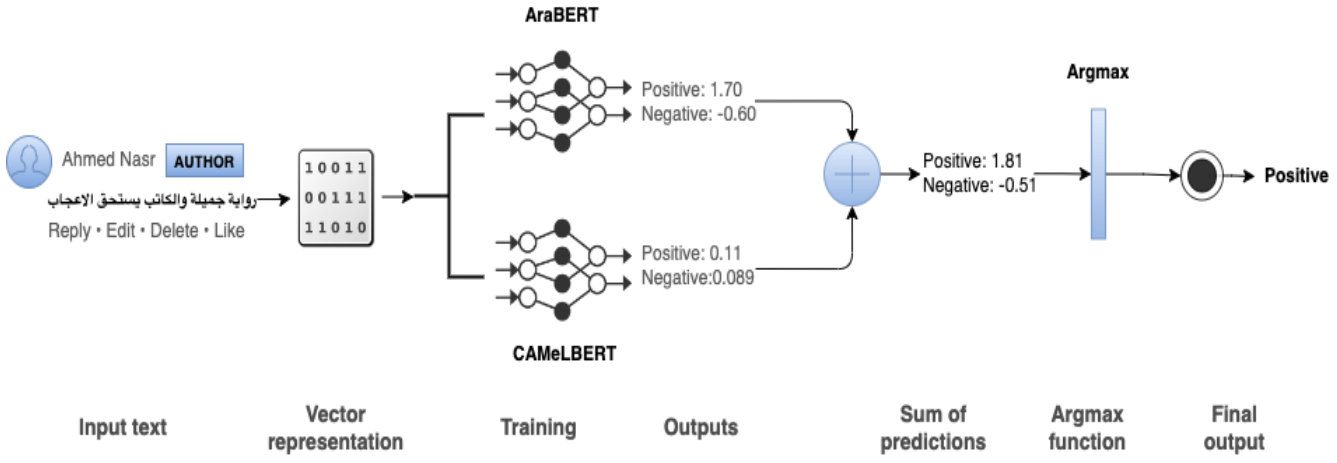


Fig. 2. The Architecture of the Proposed Ensemble Model (SUM).

TABLE II. ACCURACY RESULTS OF DIFFERENT MODELS ON THREE PUBLICALY AVAILABLE ARABIC DATASETS

			Negative class			Positive class			Accuracy	Macro-F1
			Precision	Recall	F1-score	Precision	Recall	F1-score		
Twitter Dataset		Abdulla et al. (SVM) [33]	-	-	-	-	-	-	87.2	-
	Unbalanced	Dahou et al. (CNN) [34]	-	-	-	-	-	-	85.01	-
		AraBERT	94.03	96.43	95.21	96.32	93.85	95.06	95.14	95.14
		CAMeLBERT	94.97	96.43	95.70	96.35	94.87	95.61	95.65	95.65
		Majority Voting	93.24	98.47	95.78	98.37	92.82	95.51	95.65	95.65
		SUM (ours)	95.02	96.22	96.22	97.37	94.87	96.10	**96.16**	**96.16**

	Balanced	Dahou et al. (CNN) [34]	-	-	-	-	-	-	86.3	-
		AraBERT	95.52	95.52	96.00	95.14	96.17	95.65	95.83	95.83
		CAMeLBERT	96.30	90.55	93.33	90.26	96.17	93.12	93.23	93.23
		Majority Voting	94.58	95.52	95.05	95.03	93.99	94.51	94.78	94.78
		SUM (ours)	97.87	91.54	94.60	91.33	97.81	94.46	94.53	94.53
Gold Dataset		Refaee and Rieser (SVM) [35]	-	-	-	-	-	87.74	-	
	Unbalanced	Dahou et al. (CNN) [34]	-	-	-	-	-	75.8	-	
		AraBERT	94.72	87.99	91.23	73.51	87.18	79.77	87.77	85.50
		CAMeLBERT	94.63	90.69	92.62	78.03	86.54	82.07	89.54	87.34
		Majority Voting	93.20	94.12	93.66	84.21	82.05	83.12	**90.78**	**88.39**
		SUM (ours)	94.36	90.20	92.23	77.01	85.90	81.21	89.01	86.72
	Balanced	Dahou et al. (CNN) [34]	-	-	-	-	-	-	73.8	-
		AraBERT	85.80	83.73	84.76	85.71	87.57	86.63	85.75	86.63
		CAMeLBERT	86.83	87.35	87.09	88.59	88.11	88.35	87.75	87.72
		Majority Voting	82.97	90.96	86.78	91.12	83.24	87.01	86.89	86.89
		SUM (ours)	87.95	87.95	87.69	89.13	88.65	88.89	88.32	88.29
ASTD Dataset	Unbalanced	Dahou et al. (CNN) [34]	-	-	-	-	-	-	79.07	-
		AraBERT	93.38	85.30	89.16	71.67	86.00	78.18	85.51	83.67
		CAMeLBERT	91.47	89.63	90.54	77.07	80.67	78.83	86.92	84.68
		Majority Voting	90.37	91.93	91.14	80.56	77.33	78.91	87.53	85.03
		SUM (ours)	92.42	87.90	90.10	74.85	83.33	78.86	86.52	84.48
	Balanced	Dahou et al. (CNN) [34]	-	-	-	-	-	-	75.9	-
		AraBERT	86.90	83.44	85.14	85.71	88.76	87.21	86.25	86.17
		CAMeLBERT	87.76	85.43	86.58	87.28	89.35	88.30	87.50	87.44
		Majority Voting	83.95	90.07	86.90	90.51	84.62	87.46	87.19	87.18
		SUM (ours)	89.80	87.42	88.59	89.02	91.12	90.06	**89.38**	**89.32**

IV. Experiments

In this section, we discuss the implemented models and their results. Two transformer language models are experimented namely CAMeLBERT and AraBERT, an ensemble of these two models, as well as majority voting ensemble model.

A. Dataset

For experimentation purposes, in this work we consider four datasets of different sizes and sources. The first dataset is Twitter dataset collected by Abdulla et al,. [33] composed of about 2000 tweets, written in MSA and Jordanian dialect. And consisted of 958 negative tweets and 993 positive tweets. The second dataset is Arabic Gold-Standard Twitter dataset collected by Refaee and Rieser [35] composed of 6512 tweets, divided in three classes: Negative, neutral, and positive. The negative class contains 1941 tweets, the neutral class contains 3694, and the positive class contains 876 tweets. In this study we perform a binary classification, thus only Positive and negative classes are utilized. The third dataset is Arabic sentiment tweets dataset (ASTD) proposed by [19] which contains 10006 tweets written in MSA and Arabic dialect. Tweets are labeled as one of four classes: negative, positive, neutral, and objective. As this study focuses on binary classification only positive and negative tweets are considered. After data preprocessing, we are left with 1684 negative tweets and 799 positive tweets. The fourth dataset is a dataset that we have proposed in a previous work [36] which is consisted of 1299 Modern Standard Arabic books reviews with a balance between positive and negative reviews. Reviews are collected from Goodreads website and annotated manually. After data collection, each given text is decomposed into tokens. Then, Arabic stop words are filtered out as they do not hold any information. The next step is normalization, where elongation, hamza, and diacritics are removed. Finally, all emoticons and emojis are deleted based on a preselected list of the most commonly used emoticons and emojis.

B. Results

To investigate the effectiveness of the proposed ensemble model three models have been implemented, including two transformer language-based models: AraBERT and CAMeLBERT and majority voting model. All models are trained on the same training set, which represents 80% of the whole dataset, and tested on the same testing set composed of the 20% remaining data. For each of the four datasets the models are trained and tested on both balanced and unbalanced datasets. Performance results of the proposed ensemble method are compared with stand-alone models and majority voting ensemble model. The models are evaluated using accuracy, F1-score, precision and recall metrics. The mathematical formulas of each of the used metrics is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Where, TP, FP, TN, and FN refer to "True Positive", "False Positive", "True Negative", and "False Negative" respectively.

Table II shows the performance of each model on balanced and unbalanced datasets compared to existing models. As can be seen, it is clear that ensemble models provide remarkable improvement over baseline models. Majority voting model has achieved the best accuracy score on unbalanced Gold dataset with an accuracy score of 90.78% followed by our ensemble model with 89.01%, whereas on Twitter, and ASTD datasets the best accuracy score is achieved by our proposed ensemble model SUM. On Twitter dataset SUM model achieved an accuracy score of 96.16% with unbalanced dataset followed by majority voting model and CAMeLBERT model with the same accuracy score of 95.65%. As for ASTD dataset our proposed model outperformed all other models by achieving the first best accuracy score of 89.38% on balanced dataset, the second-best accuracy score is achieved by majority voting model with 87.53% on unbalanced dataset.

The results of all proposed models on our dataset are shown on Table III. AraBERT model has achieved better results than CAMeLBERT in terms of accuracy and F1-score. On the other hand, the performance of ensemble models varies from one model to another. Compared to the proposed ensemble model, majority voting model has failed to improve the performance. It has achieved an accuracy of 94.98% against an accuracy of 95.75% achieved by AraBERT. Whereas, our proposed ensemble model has reached the best results in terms of accuracy and F1-score. The mediocre performance of majority voting may be explained by the size of the dataset and the number of combined models.

TABLE III. COMPARISON OF DIFFERENT MEASURES OF PERFORMANCE ON OUR DATASET

Model	Accuracy	F1-score	Recall	Precision
AraBERT	95.75	95.72	96.09	95.35
CAMeLBERT	92.66	92.66	93.75	91.60
Majority Voting	94.98	94.78	92.19	97.52
SUM (ours)	**96.53**	**96.50**	**96.88**	96.12

In summary, the best results have been achieved by our proposed ensemble model on balanced datasets. Thus, it is obvious from the conducted comparative experiments that training models on balanced data can improve classification performance, it can help models to learn better and achieve better accuracy results.

V. CONCLUSION

In this work, we have implemented an ensemble model based on two transformer language models, namely AraBERT and CAMeLBERT. The proposed ensemble model was evaluated on top of our balanced dataset composed of modern standard Arabic book reviews. In addition, to investigate more the performance of our proposed model it has been trained on top of three other datasets namely Twitter dataset, Gold dataset and ASTD dataset. Compared to majority voting and the two stand-alone transformer-based models, our proposed ensemble model has achieved the highest score of accuracy and F1 metrics on all datasets. In this paper, we have proposed a domain-independent model, the proposed ensemble model has achieved state-of-the-art on several datasets of different sources and domains. Thus, researchers can adopt our proposed model to address sentiment analysis in Arabic regardless of data type (MSA/Dialect) and domain. To continue working towards improving the model's performance, for future work, we plan to experiment more transformer models, combine multiple models and evaluate all possible combinations to determine the optimized model. Finally, we will be considering increasing the size of our training set as accuracy increases with the size of training data.

REFERENCES

[1] H. Al-Rubaiee, R. Qiu, and D. Li, "Identifying Mubasher software products through sentiment analysis of Arabic tweets," 2016 Int. Conf. Ind. Informatics Comput. Syst., pp. 1–6, 2016.

[2] K. M. Alomari, H. M. Elsherif, and K. Shaalan, "Arabic tweets sentimental analysis using machine learning," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, vol. 10350 LNCS, pp. 602–610, doi: 10.1007/978-3-319-60042-0_66.

[3] R. Duwairi, M. Faqeeh, M. Wardat, and A. Alrabadi, "Sentiment analysis for Arabizi text," 2016, pp. 127–132, doi: 10.1109/IACS.2016.7476098.

[4] A. Albayati and A. Al-Araji, "Arabic Sentiment Analysis (ASA) Using Deep Learning Approach," Univ. Baghdad Eng. J., vol. 26, pp. 85–93, 2020, doi: 10.31026/j.eng.2020.06.07.

[5] A. Soufan, "Deep learning for sentiment analysis of Arabic text," 2019, doi: 10.1145/3333165.3333185.

[6] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-Based Sentiment Analysis using BERT," 2019.

[7] U. Naseem, M. Khushi, V. Reddy, S. Rajendran, I. Razzak, and J. Kim, "BioALBERT: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition," in 2021 International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1–7, doi: 10.1109/IJCNN52387.2021.9533884.

[8] W. Yoon, J. Lee, D. Kim, M. Jeong, and J. Kang, "Pre-trained Language Model for Biomedical Question Answering BT - Machine Learning and Knowledge Discovery in Databases," 2020, pp. 727–740.

[9] A. S. Alammary, "BERT Models for Arabic Text Classification: A Systematic Review," Appl. Sci., vol. 12, no. 11, p. 5720, 2022.

[10] S. Bilal, "A Linguistic System for Predicting Sentiment in Arabic Tweets," in 2021 3rd International Conference on Natural Language Processing (ICNLP), 2021, pp. 134–138, doi: 10.1109/ICNLP52887.2021.00028.

[11] A. Wadhawan, "Arabert and farasa segmentation based approach for sarcasm and sentiment detection in arabic tweets," arXiv Prepr. arXiv2103.01679, 2021.

[12] A. Abuzayed and H. Al-Khalifa, "Sarcasm and Sentiment Detection In {A}rabic Tweets Using {BERT}-based Models and Data Augmentation," in Proceedings of the Sixth Arabic Natural Language Processing Workshop, Apr. 2021, pp. 312–317, [Online]. Available: https://aclanthology.org/2021.wanlp-1.38.

[13] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," 2021, doi: 10.48550/ARXIV.2101.01785.

[14] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-Training BERT on Arabic Tweets: Practical Considerations," 2021.

[15] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, May 2020, pp. 9–15, [Online]. Available: https://aclanthology.org/2020.osact-1.2.

[16] W. Lan, Y. Chen, W. Xu, and A. Ritter, "An Empirical Study of Pre-trained Transformers for {A}rabic Information Extraction," in Proceedings of the 2020 Conference on Empirical Methods in Natural

Language Processing (EMNLP), Nov. 2020, pp. 4727–4734, doi: 10.18653/v1/2020.emnlp-main.382.

[17] A. Safaya, M. Abdullatif, and D. Yuret, "{KUISAIL} at {S}em{E}val-2020 Task 12: {BERT}-{CNN} for Offensive Speech Identification in Social Media," in Proceedings of the Fourteenth Workshop on Semantic Evaluation, Dec. 2020, pp. 2054–2059, doi: 10.18653/v1/2020.semeval-1.271.

[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, Jun. 2019, vol. 1, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[19] M. Heikal, M. Torki, and N. El-Makky, "Sentiment Analysis of Arabic Tweets using Deep Learning," Procedia Comput. Sci., vol. 142, pp. 114–122, 2018, doi: 10.1016/j.procs.2018.10.466.

[20] M. Nabil, M. Aly, and A. F. Atiya, "ASTD: Arabic sentiment tweets dataset," in Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, Sep. 2015, pp. 2515–2519, doi: 10.18653/v1/d15-1299.

[21] S. R. El-Beltagy, M. El Kalamawy, and A. B. Soliman, "NileTMRG at SemEval-2017 Task 4: Arabic Sentiment Analysis," in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Aug. 2017, pp. 790–795, doi: 10.18653/v1/S17-2133.

[22] S. R. El-Beltagy, "NileULex: A phrase and word level sentiment lexicon for Egyptian and modern standard Arabic," in Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, 2016, pp. 2900–2905.

[23] A. Vaswani et al., "Attention is All you Need," in Advances in Neural Information Processing Systems, 2017, vol. 30, [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[24] K. Clark, M.-T. Luong, Q. V Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." arXiv, 2020, doi: 10.48550/ARXIV.2003.10555.

[25] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv, 2019, doi: 10.48550/ARXIV.1907.11692.

[26] A. Radford and K. Narasimhan, "Improving Language Understanding by Generative Pre-Training," 2018.

[27] W. Antoun, F. Baly, and H. Hajj, "{A}ra{BERT}: Transformer-based Model for {A}rabic Language Understanding," in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, May 2020, pp. 9–15, [Online]. Available: https://aclanthology.org/2020.osact-1.2.

[28] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The Interplay of Variant, Size, and Task Type in {A}rabic Pre-trained Language Models," in Proceedings of the Sixth Arabic Natural Language Processing Workshop, Apr. 2021, pp. 92–104, [Online]. Available: https://aclanthology.org/2021.wanlp-1.10.

[29] A. Safaya, "Arabic-ALBERT." Zenodo, Aug. 2020, doi: 10.5281/zenodo.4718724.

[30] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul. 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.

[31] I. A. El-khair, "1.5 billion words Arabic Corpus," ArXiv, vol. abs/1611.0, 2016, [Online]. Available: http://arxiv.org/abs/1611.04033.

[32] I. Zeroual, D. Goldhahn, T. Eckart, and A. Lakhouaja, "{OSIAN}: Open Source International {A}rabic News Corpus - Preparation and Integration into the {CLARIN}-infrastructure," in Proceedings of the Fourth Arabic Natural Language Processing Workshop, Aug. 2019, pp. 175–182, doi: 10.18653/v1/W19-4619.

[33] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, AEECT 2013, 2013, pp. 1–6, doi: 10.1109/AEECT.2013.6716448.

[34] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud, and P. Duan, "Word embeddings and convolutional neural network for Arabic sentiment classification," in COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers, Dec. 2016, pp. 2418–2427, [Online]. Available: https://www.aclweb.org/anthology/C16-1228.

[35] E. Refaee and V. Rieser, "An Arabic twitter corpus for subjectivity and sentiment analysis," in Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, 2014, pp. 2268–2273.

[36] I. El Karfi, S. El Fkihi, and R. Faizi, "A spectral clustering-based approach for sentiment classification in modern standard Arabic," in MCCSIS 2018 - Multi Conference on Computer Science and Information Systems; Proceedings of the International Conferences on Big Data Analytics, Data Mining and Computational Intelligence 2018, Theory and Practice in Modern Computing 2018 and Connected Sma, 2018, pp. 59–65, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054167622%5C&partnerID=40%5C&md5=101982324dd788664f052d2919012106.

Grover's Algorithm for Data Lake Optimization Queries

Mohamed CHERRADI, Anass EL HADDADI
Data Science and Competitive Intelligence Team (DSCI)
ENSAH, Abdelmalek Essaadi University
Tetouan, Morocco

***Abstract*—Now-a-days, the use of No-SQL databases is one of the potential options for storing and processing big data lakes. However, searching for large data in No-SQL databases is a complex and time-consuming task. Further, information retrieval from big data management suffers in terms of execution time. To reduce the execution time during the search process, we propose a fast and suitable approach based on the quantum Grover algorithm, which represents one of the best-known approaches for searching in an unstructured database and resolves the unsorted search query in O ($\sqrt{n}$) time complexity. To assess our proposal, a comparative study with linear and binary search algorithms was conducted to prove the effectiveness of Grover's algorithms. Then, we perform extensive experiment evaluations based on ibm_qasm_simulator for searching one item out of eight using Grover's search algorithm based on three qubits. The experiments outcomes revealed encouraging results, with an accuracy of 0.948, well in accordance with the theoretical result. Moreover, a discussion of the sensitivity of Grover's algorithm through different iterations was carried out. Then, exceeding the optimal number of iterations *round* ($\frac{\pi}{4}\sqrt{N}$), induces low accuracy of the marked state. Furthermore, the incorrect selection of this parameter can outline the solution.**

Keywords*—Big data; data management; information retrieval; quantum computing

I. INTRODUCTION

In the last decades, database management systems have occupied a significant area in IT due to their efficiency in managing massive amounts of heterogeneous datasets. Indeed, the investigation of database research leads to the evolution of special concepts, processes, and algorithms. However, big data lake, recent big news, depicts a recommended solution for dealing with heterogeneous datasets in any format, structured, semi-structured, or unstructured. Thus, numerous contributions, such as No-SQL databases, have been offered for the optimization of processing times on the Big Data Lake [1] [2] [3]. Faced with this challenge, this paper aims to investigate the data lake optimization queries through an efficient and powerful approach based on the Grover algorithm. As the volume of data generated grows, the requirements for a large data processing supercomputer has attracted increasing interest due to their various applications. Therefore, using quantum computers as very fast calculators represents one of the hot topics for accelerating big data processing. It allows us to drastically reduce the execution time when searching for data in a large space. The Grover algorithm was introduced as one of the most beneficial algorithms for data lakes.

Although various classical data retrieval methods have been proposed, most of them remain heavy in query execution for a big data space, which is characterized by volume, variety, and veracity, among other v-properties. This constitutes a significant issue, as several applications are defined in large-scale environments with heterogeneous data in which the majority of the data is unstructured, almost 80%. Furthermore, searching in an unstructured database using quantum algorithms is one of the most widely used techniques to speed up classical search algorithms. It allows finding a more generic research solution to a very wide range of problems [4]. The search time required for a database depends on the size of the database and the quantum hardware. Therefore, it turns out that it is necessary to analyze the design of the quantum circuit.

To address the execution latency issue when searching in a challenging big data space. In this paper, we propose to investigate data lake optimization queries using an efficient and powerful approach based on the Grover algorithm, which is the fastest quantum algorithm for searching an unsorted database with a quandratic complexity of O ($\sqrt{(N)}$) time, as opposed to classical algorithms with a linear complexity of $O(N)$ time. Roughly speaking, a standard analogy for Grover's algorithm is to look up the name of a person in a phone book who only knows their phone number. The phone book remains an unsorted database, and a classical search algorithm appears tedious. On average, this would take N requests, or $N/2$ in the worst case, depending on the position of the desired element, with N denoting the number of entries in the telephone annuaire. Yet, if the correlation between phone name and number is encoded or embedded with quantum bits, the search phone number is reduced approximately to $\sqrt{(N)}$ requests. Thus, quantum computing is a fast-evolving domain, and it is reaching significant accelerations compared to classical algorithms [6] [7] [8] [9]. Considering the speed at which data is growing every day, it is necessary to think of powerful algorithms with the ability to process data quickly and efficiently. Based on this, the principal motivation of this research article is to propose the quantum design of the Grover algorithm and benefit from its speed-up to efficiently manage and extract the hidden relevant information from the heterogeneous data lake. Our proposed, implement IBM Quantum Composer to build the Grover quantum circuit. Indeed, IBM provides multiple quantum computers to the public through its IBM cloud service, accessible via the

application programming interface such as Qiskit [5]. The experiments prove Grover's algorithm as one of the most beneficial algorithms for data lakes.

The remainder of this article is organized as follows: Section II provides the necessary background for readers to fully understand our article. Section III presents the different stages of Grover's algorithm. Moreover, results and discussions are examined in Section IV. Finally, we conclude with a summary and some perspectives in Section V.

II. Related Work

In this section, we review some preliminary and necessary background information needed for the readers to fully understand the rest of our article. We start by examining the data lake concept as a storage space for heterogeneous data sources. Thus, we will give an overview of all the concepts related to quantum computing.

A. Data Lake Concept

In the last decades, the amount of data produced every day is absolutely horrible. So-called big data refers to the exponential growth of massive data. In this context, J. Dixon [10] introduced the data lakes concept to address the challenges and issues induced by big data. Among one of the principal issues studied in the literature is metadata management, proposed with the objective of avoiding the transformation of data lakes into data swamps, i.e., useless data [11] [12] [13] [14]. Thus, data lakes have evolved into data management solutions capable of meeting big data needs and producing a high level of advanced data analysis. They accept various data sources and can accommodate a resilient ecosystem for making creative, data-driven business determinations. Also, Data Lake has a data-centric approach, which refers to an architecture in which data is the primary and permanent asset. Therefore, the data lake has developed as a strong and adaptable concept better suited to data analytics, allowing enterprises to take advantage of this complicated data and generate new commercial industrial activities. While traditional ETL is used in data warehouses to prepare data for integration into a structured relational database, ELT (Extract, Load, and Transform) paradigms are used in data lakes to process unstructured data [15] [16] [17]. Data is loaded into the lake "as-is", with no data transformation. This makes it easier to set up jobs because all that is required is a declaration of the origin and destination locations. As a result, one can reduce the time spent on the data transformation phase, which is considered the most expensive stage in any data project, accounting for over 60% of the total time spent on the project.

Since 2016, the contributions of data lakes in both industry and the academic community have been growing. But most of the data lake proposals are abstract and depend on a specific use case. In our case, we will try to project Grover's algorithm into the data lake as being an unstructured data search space. Since this algorithm applies to unstructured data, it adapts perfectly to the data lake to find crucial information stored in the lake.

B. Quantum Computing

Today's conventional computers are marked with "classical bits" (cbits), which are the basic units of data. With one bit, it takes either the value 0 or 1. Yet, this type of computer faces a limit when challenged with a multivariate problem. In this case, each calculation is a unique path to a unique result. Furthermore, classical computers are less efficient in terms of computation compared to quantum computers due to the limits of classical physics principles, which constitute the core of classical computer components [18] [19]. Thereby, due to recent hardware advancements, quantum computing is a rapidly evolving research field. The principles of quantum mechanics enable quantum computers to solve certain classes of problems very quickly compared to classical computers. Such as factorization and searching databases [21] [22] [23]. Moreover, quantum computers are classified as supercomputers because they exploit the strengths of quantum mechanics, including the quantum superposition principle and entanglement [20]. The superposition principle reflects the possibility of considering a quantum system to be in multiple states at the same time. While quantum entanglement defines the correlation between two (or more) quantum particles even though they are distantly separated.

Following the classical nature of the binary bit, the qubit tries to design a superposition of the states $|0\rangle$ and $|1\rangle$. Since a quantum system can be prepared in a superposition state, the quantum computer can perform 2^n calculations in a single physical step, where n represents the number of qubits used during this process [24]. Furthermore, the quantum computer can execute jobs in exponentially fewer steps than a conventional computer. A qubit can be expressed as a unit vector in a complex vector space, C^2. Constantly written in the form of ket and bra, which corresponds to the notation of Dirac [25]. Hence, the qubit at state zero is written as $|0\rangle$ and the qubit at state one is written as $|1\rangle$. $|0\rangle$ and $|1\rangle$ represent the basis vectors in the complex vector space of quantum states. A Bloch sphere, observed in Fig. 1, is used as a geometric representation of the qubit. The state $|1\rangle$ is represented by the south pole of the sphere, while the state $|0\rangle$ is represented by the north pole. A state $|\psi\rangle = \alpha_0 |0\rangle + \alpha_1 |1\rangle$ is defined as the angle point (θ, ϕ), where α_0 and α_1 validate the normalization condition, i.e. $\alpha_0^{\ 2} + \alpha_1^{\ 2} = 1$. Thus, it is written in the geometric form with $\alpha_0 = cos\ \theta/2$ and $\alpha_1 = e^{i\phi} sin\ \theta/2$. Yet, the Bloch sphere can be very useful as a geometric representation to visualize the quantum state and its transformation.

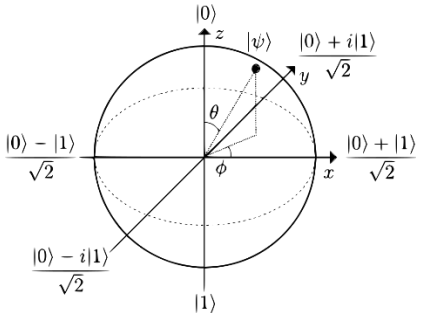


Fig. 1. Bloch Sphere [19].

Further, quantum computing, as one of the rapidly emerging research topics, has fundamentally altered the computing world. Quantum software development continues to be one of the most active investigative study fields [26]. This implies the proposal of new algorithms that adapt to a specific type of new information processing technology. Therefore, quantum computing fascinates the scientific community of researchers because it shows the computing power of big data in a reduced time. Thereby, quantum computing could stimulate scientific progress by leveraging quantum mechanical theories.

However, before we can assess the quantum computer's advantages, several restrictions must be addressed. The most famous one is the decoherence phenomenon, which is the major obstacle. Indeed, to calculate much faster than a conventional computer, the quantum computer uses superposition and entanglement of states that are significantly more sensitive to the environment than classical states [27]. The more qubits you add to a system, the more parallel operations you will increase. Then, since the environment interacts with qubits, quantum measurement uncontrollably changes quantum states. This is called decoherence and is caused by a variety of factors in the environment, including changes in magnetic and electric fields, radiation from nearby hot objects, or uncontrolled interactions between qubits, among others. Subsequently, decoherence affects the state of superposition and disrupts quantum information processing. It is the biggest barrier to the development of quantum technology. Furthermore, it is crucial to examine quantum computing technologies and algorithms.

1) Technologies of quantum computing: The quantum computing field has seen tremendous technological *advancement* over the past decades. Nonetheless, the state of the art related to quantum computing technologies [28] is provided by web giants, such as IBM, Google, Intel, and Microsoft. IBM is one of the major corporations that has made significant investments in quantum computing [29]. At the time of writing this article, IBM had almost 12 simulators, which had up to 5000 qubits, corresponding to a simulator called "simulator stabilizer". Thus, there is a simulator with only one qubit, which corresponds to a simulator called "ibmq armonk". IBM's simulators employ IBM QISKit, a highly handy python library, to process asynchronously run jobs [30] [31]. Qiskit is an open-source framework for quantum computing. It provides the necessary tools that can be used to create and manipulate quantum programs and run them on prototype quantum devices on the IBM Q Experience or simulators on a local computer. Furthermore, once a job process is completed, the user receives the results in the form of the job run time (seconds) and the measurement of each state. Moreover, IBM provides multiple quantum computers to the public through its IBM cloud service. The ibmqx5 is a 16-qubit superconductivity-based quantum computer, ready through an Application Programming Interface (Python-API) called QISKit.

2) Quantum algorithms: In the quantum computing era, a quantum algorithm is a quantum computation solution that works on a practical quantum model [32]. It is often designed as a quantum circuit. Moreover, a classical (or non-quantum) algorithm is a sequence of instructions for solving a problem. One of the most well-known classical search algorithms is that of sequential and interval search.

a) Sequential search: One of the most basic and simplest search algorithms that fall under the category of searches is linear. This type of algorithm works sequentially (without jumping) through a list by comparing each element with the value we want to find [33]. In the worst case, the time complexity corresponds to the order of N, indicated as O(N), where N represents the number of elements in the list. This algorithm has the advantage of not requiring the list to be sorted because it works regardless of the order in which the list's elements appear. However, finding the element you're seeking takes a long time. As long as the number of elements in the list is large, the algorithm takes a lot of time.

b) Interval search: One of the frequently used implementations in interval search is the binary search. In fact, the search space must be ordered. Furthermore, this sort of algorithm divides the collection of elements that make up the search space into intervals [34], such that if the search value is smaller than the value in the middle of the interval, in this case, the search is not performed only at a level of less than half the interval. Otherwise, the search is carried out at the upper level. This process is repeated until the element marked is found in logarithmic time.

III. GROVER'S ALGORITHM

The study of quantum algorithms has recently been one of the most difficult scientific issues that has radically transformed the way people think about computers. Indeed, quantum computing remains a part of worldwide reality, and its advancement cannot be overlooked. Working on this research topic was also the goal of former researchers [35]. The principles of quantum physics, like, for example, the use of superconducting quantum processors, have a major peculiarity [36]. Exercising superconducting quantum circuit technology enables the researchers to contribute a list of contributions related to the quantum algorithms. In 2016, IBM introduced the Quantum Experience program, which provides a set of online quantum simulators that allow anyone interested to execute their quantum circuit [37]. The IBM Quantum Experience handbook gives users a hands-on experience with all of the criteria for building a quantum circuit that solves a specific problem perfectly.

Grover's algorithm offers a quick search through a mass of unstructured data to find the desired information. It has proven a significant speedup compared to the classical algorithm and produced a promising result, motivating extensive investigation into the viability of applying Grover's algorithm to a variety of domains. In this paper, we examine a search space of size N with no prior knowledge of how the data will be presented. This problem has a polynomial complexity with classical solutions, whereas the quantum search algorithm has a quadratic complexity $O(\sqrt{(N)})$ [38]. Through this paper, we have proposed an overview in algorithmic form, summarizing

the different stages of Grover's algorithm. As shown in algorithmic prototype 1.

Algorithm 1 Grover's Algorithm for data lake

Input: Heterogeneous datasets form a data lake $DL = \{x_0, x_1, x_{N-1}\}$

Output: Get the index of the marked element $x^* \in N$

Step 1: The quantum register's initialization:

Set the state of all qubits $x \otimes^n$ to the state $|0\rangle$ and set the oracle qubit to $|1\rangle$ state : $|\psi_0\rangle = |0\rangle \otimes^n |1\rangle$

Step 2: Deploy the register in a distributed uniform superposition:

Apply the Hadamard gate H:

$$|\psi_1\rangle = \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} |i\rangle \otimes \frac{|0\rangle - |1\rangle}{\sqrt{2}}$$

Step 3: Repeat Grover's iterations:

for round($\frac{\pi}{4}\sqrt{N}$) times; do

a- Apply the oracle:

$|x\rangle \rightarrow (-1)^{f(x)} |x\rangle$

b- Execute Grover operator (reflection about the mean)

1. Apply $H \otimes^n$
2. Conditionally shift phase
3. Apply $H \otimes^n$

End for

Step 4: Quantum register measurement

Finding the index of the target element in a list of $N = 2^n$ entries is the problem of searching in an unordered list. With n denoting the number of qubits and N denoting the list's length. Moreover, an unstructured search is commonly expressed as a database search query in which we want to find an item that meets a set of criteria specified in the query. We refer to the problem search as "unstructured" because we have no control over how the data is organized in the database. If we have an ordered database, we can use a binary search to find the predicted element in logarithmic time. However, if we don't know the sequence of the database items, the task remains difficult to complete in terms of execution, and we can't get better results with the conventional approach. If there is no indication of where the desired item might be found, in this case, any classical algorithm must examine each element individually. Furthermore, the number of tries required to find the sought item equals the number of items in the list. As we can see, using quantum mechanics principles, only $O(\sqrt{(N)})$ attempts are required. To meet this requirement, Grover's algorithm uses two registers, the first one linked with quantum qubits, in which we shall create a superposition of all 2^n basis states $\{|0\rangle, ..., |2N-1\rangle\}$. This can be done by applying the Hadamard gate to all the initial qubits. While the second register is linked to classical bits to persist the measurement results, it takes either the value of *0* or *1*. For the sake of precision, we describe the different stages of Grover's algorithm:

A. Initialization

The Grover algorithm starts by initializing the qubits in the state $|0\rangle$ by performing a uniform superposition of all basic inputs. A Hadamard quantum gate, given by the matrix (1), is implemented to create a superposition of the set of quantum states [39].

$$H=\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix} \quad (1)$$

By applying the Hadamard gate to state $|0\rangle$, we obtain the following state.

$$\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}1\\ 0\end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\ 1\end{bmatrix}$$

Which has mapped:

$$|0\rangle \rightarrow \frac{|0\rangle + |1\rangle}{\sqrt{2}}$$

If we instead initialize the qubit to $|1\rangle$ and apply a Hadamard gate:

$$\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\ -1\end{bmatrix}$$

Which has mapped:

$$|1\rangle \rightarrow \frac{|0\rangle - |1\rangle}{\sqrt{2}}$$

As a result, to generalize the Hadamard gate application to the initial state $|0\rangle$, we obtain the following formula:

$$H^{\otimes n} |0\rangle = \sum_{i=0}^{N-1} \alpha_i |i\rangle$$

Where α_i represents the amplitude probability. Indeed, all quantum states have the same amplitude, i.e., $\alpha_i = \frac{1}{\sqrt{N}}$.

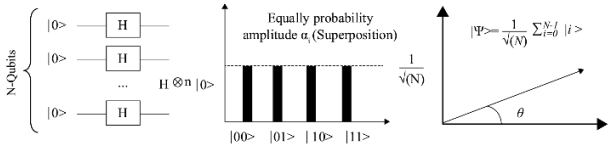


Fig. 2. The Grover Algorithm's Initialization Step.

As illustrated in Fig. 2, for the case of the number of qubits $N = 4$. Therefore, the number of possible states corresponds to $N = 2^n = 2^2 = 4$. Each state is associated with equiprobable amplitudes, $\alpha_i = \frac{1}{\sqrt{4}} = \frac{1}{2}$.

B. Oracle

After having initialized the circuit with the Hadamard gate to create a superposition of quantum states, Grover's algorithm will proceed through its first iteration, which corresponds to what is known as the quantum oracle. The oracle, also known as a "black-box" function, modifies the quantum state of the item's index we're seeking [40] [41]. The change of the quantum state by the oracle Grover was performed without transforming it into a classical state. If the system is located in the right state, then the oracle will turn the phase by the angle π. Otherwise, no action will be taken. The function f corresponds to the oracle expressed as follows:

$$f(x)=\begin{cases}1 & if\ x\ is\ the\ correct\ state\ ,\\ 0 & Otherwise\ .\end{cases} \quad (2)$$

The quantum circuit implements the function *f* described by the unitary operator denoted by *O*.

$$O|\psi\rangle = \sum_i \propto_i (-1)^{f(i)} |i\rangle \quad (3)$$

The function *f* verifies the searched item by transforming the sign of its probability amplitude if *f(x) = 1*. Otherwise, nothing happens. To illustrate the operation of the oracle, we take an example of two qubits. To create a superposition state, we apply the Hadamard gate.

$$|\phi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |01\rangle + |10\rangle + |11\rangle) \quad (4)$$

Suppose that the item we're looking for is the index marked by $|i\rangle = |i^*\rangle = |10\rangle$. By applying the oracle to the state $|\phi\rangle$, we get the state $|10\rangle$ signed by a phase of factor -1. Grover's algorithm oracle step is depicted in Fig. 3.

$$|\phi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |01\rangle - |10\rangle + |11\rangle) \quad (5)$$

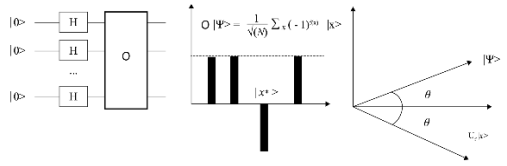


Fig. 3. The Grover Algorithm's Oracle Step.

C. Amplification

The amplification step performs a reflection around the average of the amplitudes. It flips the target state by increasing its amplitude probability and decreasing other states. Yet, this step can be implemented by a combination of the following gates: *HRH*. Here, *H* designates the Hadamard gate, and and *R* designates a phase shift transform [21]. Fig. 4 depicts Grover's algorithm amplification step.

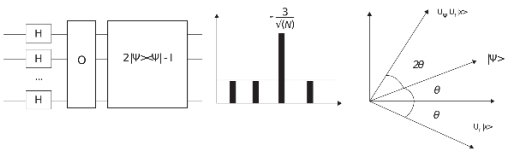


Fig. 4. The Grover Algorithm's Amplification Step.

D. Measurement

The measurement of each qubit in a quantum circuit is performed as the last step of the calculation to produce an output in the classical bit [42]. Indeed, we cannot say that a qubit has an actual value, but rather that it contains a probability of being found in a particular state when measured. Moreover, the measurement step is necessary to derive a result from the quantum state computation. The quantum gate associated with this step (the measurement gate) represents the only non-reversible quantum gate.

IV. Results and Discussions

Suppose we want to find the name of a well-described article with a set of metadata (Fig. 5). Each article that exists is indexed by an integer belonging to segment $\{0, \ldots, N-1\}$. Given that, we're looking for an article with an index $i = I^*$. The articles are not ordered, and we need to get a particular record from the list of articles. If we use the classical algorithm, we may be lucky and find the article we are looking for in the first index, i.e., $i = I^* = 0$, or we may not find the article until the last index $i = I^* = N-1$. Furthermore, for a search in the unstructured database, an average of approximately *N/2* (or *N* in the case where we found the article in the last index) of queries is required to find the article that adapts to the search criteria. It is important to point out that in the case of a uniform probability, we have a probability of *1/N* of finding an article among the *N* articles. Then, we can prove the average number of queries needed to find the right article, according to the equations below.

$$\sum_{i=1}^{N} i\frac{1}{N} = \frac{1}{N}\sum_{i=1}^{N} i,$$

$$\sum_{i=1}^{N} i = \frac{N(N+1)}{2},$$

$$\frac{1}{N}\sum_{i=1}^{N} i = \frac{1}{N}\frac{N(N+1)}{2} = \frac{(N+1)}{2} \approx \frac{N}{2} \quad (6)$$

Grover's algorithm bettered the classical search method by a quadratic speedup. The computer scientist, Grover, found a quantum search algorithm that requires only $O(\sqrt{(n)})$ steps. Suppose, for example, N=1000; the classical search algorithms do 1000 iterations (or 1000/2 = 500 in the worst case) to find the search record. However, the Grover algorithm will only perform $\sqrt{(1000)}$=100 iterations. Consequently, Grover's algorithm exhibits a significant acceleration. We cannot do great than a quadratic speedup with a complexity of order $\sqrt{(N)}$. The *N* articles are numbered from 0 to *N−1*, requiring *n* qubits to represent the list of articles (with $N = 2^n$). We can represent all *N* articles using only the principle of superposition with *n* qubits. A quantum state, $|\psi\rangle$ is designed by a column vector of size $(2^n, 1)$ whose values are probability amplitudes. Each probability amplitude is associated with a well-defined article that is identified by an index *i*.

$$|\psi\rangle = \begin{bmatrix}\alpha_0\\ \vdots \\ \alpha_m\end{bmatrix} \rightarrow \alpha_0|0\rangle + \alpha_1|1\rangle + \quad (7)$$

The article of index *i* is linked to the probability amplitude α_i. As a result, the probability of finding the article we're seeking is extremely close to 1, and the amplitude of all other probabilities is close to 0.

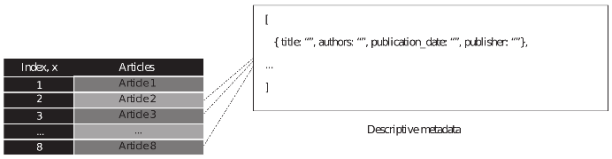


Fig. 5. Searching a List of Articles.

A. Use Case Application

Considering Data Lake (*DL*) as a database that stores heterogeneous data regardless of its format. Then, *DL*, which is made up of *N = 8* data sources, is derived from scientific

databases (Fig. 5). Each data source represents a scientific article, which is identified by a collection of descriptive metadata such as title, authors, and publication date, as well as the path where the paper is placed, which provides the paper's unique identity (*ident*). While $ident \in \{0, \ldots, N-1\}$, we want to find the article titled *X*, which contains an identifier id_k. To meet this requirement, we must first meet the prerequisite:

$f(id_k) = 1$.

Then,

$$O_f |X\rangle = -|X\rangle . \quad (8)$$

Let us express the different quantum states at each step shown in the circuit. Let's start with the first state $|\psi_0\rangle$, and end with the last state $|\psi_f\rangle$.

$$|\psi_0\rangle = |000\rangle \quad (9)$$

Subsequently, by applying the Hadamard gate, the state $|\psi\rangle$ becomes.

$$|\psi_0\rangle = H^{\otimes 3} |000\rangle = \frac{1}{\sqrt{8}} \sum_{i=0}^{7} |i\rangle = \frac{1}{2\sqrt{2}} \sum_{i=0}^{7} |i\rangle . \quad (10)$$

Consider that we are looking for the element which has the index $i = 5$. Then, $|i\rangle = |5\rangle = |101\rangle$ (Fig. 6 depicts the quantum circuit that corresponds to determining the quantum state $|101\rangle$). At this point, we need to specify the oracle operator that will be used in our use case. Indeed, when solving an NP problem, the defined oracle operator can mark the corresponding state. Therefore, the oracle operator must mark the element with the index 101 that we are looking for. Then, we have.

$$O_f |101\rangle|-\rangle = -|101\rangle|-\rangle.$$

$$O_f |i\rangle |-\rangle = |i\rangle |-\rangle \ if \ i \neq 5. \quad (11)$$

After specifying the sought state, we need to define a vector orthogonal to it, denoted by $|u\rangle$ as expressed below:

$$|u\rangle = \frac{1}{\sqrt{7}} \sum_{\substack{i=0, \\ i \neq 5}}^{7} |i\rangle, \quad (12)$$

$$|u\rangle = \frac{|000\rangle+|001\rangle+|010\rangle+|011\rangle+|100\rangle+|110\rangle+|111\rangle}{\sqrt{7}}$$

Then, we have

$$|\psi\rangle = \frac{\sqrt{7}}{\sqrt{8}} |u\rangle + \frac{1}{\sqrt{8}} |101\rangle = \frac{\sqrt{7}}{2\sqrt{2}} |u\rangle + \frac{1}{2\sqrt{2}} |101\rangle. \quad (13)$$

With this equality, one can determine the value of the angle θ as follows:

$$\theta = 2 \arccos\left(\frac{\sqrt{7}}{2\sqrt{2}}\right) \approx 41.4° \quad (14)$$

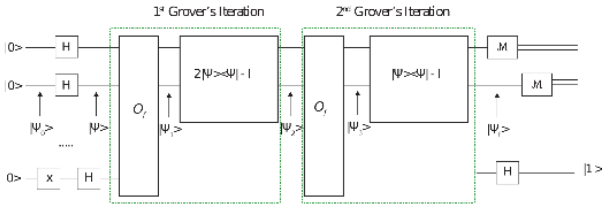


Fig. 6. Grover's Algorithm Quantum Circuit over an Unstructured Database of Eight Elements.

The next step is intended to apply the oracle operator $|\psi_1\rangle |-\rangle$. As a result, we get.

$$|\psi_1\rangle|-\rangle = O_f (|\psi 1\rangle|-\rangle) ,$$

$$= \frac{|000\rangle+|001\rangle+|010\rangle+|011\rangle+|100\rangle-|101\rangle+|110\rangle+|111\rangle}{\sqrt{7}} . \quad (15)$$

Note that quantum state $|101\rangle$ is the only one that has a minus sign. It is now suitable to rewrite $|\psi_1\rangle$ as follows:

$$|\psi_1\rangle = |\psi\rangle - \frac{1}{2\sqrt{2}} |101\rangle. \quad (16)$$

Or it can be expressed as

$$|\psi_1\rangle = \frac{\sqrt{7}}{2\sqrt{2}} |u\rangle - \frac{1}{\sqrt{2}} |101\rangle \quad (17)$$

Eq. (16), is useful for the next step's calculation. Since we are going to apply the formula $(2 |\psi\rangle \langle\psi| - I)$. The formula of Eq. (17), is useful to schematize the quantum state $|\psi_1\rangle$. Yet, the $|\psi_1\rangle$ state represents the reflection of $|\psi\rangle$ respecting the state $|u\rangle$. In the next step, we will apply the reflection again around the average.

$$|\psi_2\rangle = (2|\psi\rangle \langle\psi| - I) |\psi_1\rangle \quad (18)$$

By using Eq. (16), we get

$$|\psi_2\rangle = \frac{1}{2} |\psi\rangle + \frac{1}{\sqrt{2}} |101\rangle \quad (19)$$

Therefore, by using Eq. (13), we get

$$|\psi_2\rangle = \frac{\sqrt{7}}{4\sqrt{2}} |u\rangle + \frac{5}{4\sqrt{2}} |101\rangle \quad (20)$$

To assert that the angle between $|\psi\rangle$ and $|\psi_2\rangle$ is θ, note that

$$cos(\theta) = \langle\psi_2|\psi\rangle = \frac{1}{2} \langle\psi|\psi\rangle + \frac{1}{\sqrt{2}} \langle\psi|101\rangle = \frac{3}{4} \quad (21)$$

Which conforms with equality (14). This completes the first iteration of the Grover application designated by *G*. The second and final application of the Grover operator is similar to the first one. The next step in our examination is the analysis of the state $|\psi_3\rangle$, which is found by applying the oracle operator, as shown below:

$$|\psi_3\rangle = \frac{\sqrt{7}}{4\sqrt{2}} |u\rangle - \frac{5}{4\sqrt{2}} |101\rangle \quad (22)$$

Using Eq. (16), we get

$$|\psi_3\rangle = \frac{1}{2} |\psi\rangle - \frac{3}{2\sqrt{2}} |101\rangle \quad (23)$$

It is important to note that the state $|\psi_3\rangle$ represents the reflection of the state $|\psi_2\rangle$ with the state $|u\rangle$. Finally, the last step is to apply the inversion around the mean.

$$|\psi_f\rangle = 2(\langle\psi|\psi\rangle - I) |\psi_3\rangle \quad (24)$$

Using the two equations (13) and (23), we get

$$|\psi_f\rangle = \frac{-\sqrt{7}}{8\sqrt{2}} |u\rangle + \frac{11}{8\sqrt{2}} |101\rangle \quad (25)$$

It is self-evident θ that is the angle formed by the two quantum states, $|\psi_f\rangle$ and $|\psi_2\rangle$. Note that the amplitude of the marked state $|101\rangle$ is greater than the other quantum states $|i\rangle$, with $i \neq 101 = 5$. Subsequently, measuring the state based on

the computation will project it into quantum state 101 with the following probability:

$$p = \left|\frac{11}{8\sqrt{2}}\right|^2 = \left|\frac{121}{128}\right| \approx 0.945 \quad (26)$$

Therefore, after two iterations of applying Grover's operator, the chance of getting the sought result, which corresponds to the state $|101\rangle$ achieves an accuracy of nearly 94.5%. In the rest of this use case, we will show how important it is to know the major impact of the number of iterations of the Grover algorithm on accuracy. Suppose the number of iterations is unknown in advance. In this case, we will perform additional Grover iterations as follows:

$$|\psi_5\rangle = \frac{-\sqrt{7}}{8\sqrt{2}}|u\rangle - \frac{11}{8\sqrt{2}}|101\rangle = \frac{-1}{4}|\psi\rangle - \frac{5}{4\sqrt{2}}|101\rangle \quad (27)$$

The stage of the inversion around the mean induces the state $|\psi_6\rangle$, which is represented.

$$|\psi_6\rangle = 2(\langle\psi|\psi\rangle - I)\,|\psi_5\rangle$$

$$= 2(\langle\psi|\psi\rangle - I)\,(\frac{-1}{4}|\psi\rangle - \frac{5}{4\sqrt{2}}|101\rangle)$$

$$= \frac{-7}{8}|\psi\rangle + \frac{5}{4\sqrt{2}}|101\rangle$$

$$= \frac{-7}{8}(\frac{\sqrt{7}}{2\sqrt{2}}|u\rangle + \frac{1}{2\sqrt{2}}|101\rangle) + \frac{5}{4\sqrt{2}}|101\rangle$$

$$= \frac{-7\sqrt{7}}{16\sqrt{2}}|u\rangle + \frac{13}{16\sqrt{2}}|101\rangle \quad (28)$$

The measurement of the state $|\psi 6\rangle$ turns out to be us.

$$p = \left|\frac{13}{16\sqrt{2}}\right|^2 = \left|\frac{169}{512}\right| \approx 0.336 \quad (29)$$

Now, if we perform a measurement on the other states, the corresponding probability is calculated as below:

$$p = \left|-\frac{7\sqrt{7}}{16\sqrt{2}}\right|^2 = \left|\frac{343}{512}\right| \approx 0.67 \quad (30)$$

Table I shows the performance of the Grover algorithm according to the number of iterations. We notice that the probability of finding a solution for a search space of a specified size varies according to the number of iterations.

TABLE I. PERFORMANCE MEASUREMENT OF THE DIFFERENT ITERATIONS OF GROVER'S ALGORITHM

Simulator	*No. of Grover Iterations*	*Accuracy*
ibm_qasm_simulator	1	0.78
	2	0.945
	3	0.67

Therefore, if we continue the number of Grover iterations after the optimal number of *round*$(\frac{\pi}{4}\sqrt{N})$, the probability of finding the sought state decreases while the probability of error increases more and more. In the event of exceeding the number of iterations, which in our instance is two, the accuracy decreases by a percentage of 0.275. Thus, we report the empirical implementation of Grover's quantum search algorithm on the IBM quantum simulator with three qubits. Fig. 7 illustrates well the theoretical results that we have carried out. The QISKit code for the implementation can be found on my GitHub under the link: https://github.com/cherradii/Grover_Quantum_Search_Alg.

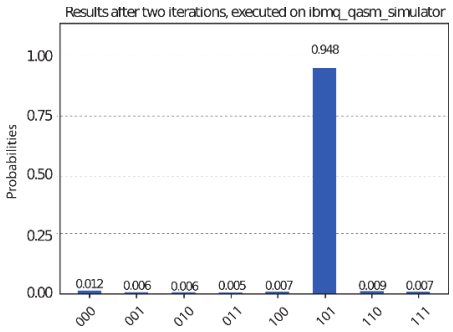


Fig. 7. Searching for Quantum State $|101\rangle$.

B. Iteration of Grover's Algorithm

Grover's algorithm is made up of a quantum subroutine named Grover's iteration, noted *G*, which is broken down into two steps:

- Apply the oracle U_f
- Apply the diffusion operator *G* on the first *n* qubits.

The iterations of Grover's algorithm are seen from a geometric point of view as a rotation in the two-dimensional space wrapped by the two vectors $|\alpha\rangle$ and $|\beta\rangle$. $|\alpha\rangle$ denotes normalized states of the sum of all targets, and $|\beta\rangle$ denotes normalized states of the sum of non-targets. The initial state $|S\rangle$ can be written as follows:

$$|S\rangle = sin\,(\theta)|\alpha\rangle + cos\,(\theta)|\beta\rangle \quad (31)$$

When looking in a search space of $N = 2^n$ items, there are *M* targets for searching ($0 \le M \le N$). Since $sin\,(\theta) = \sqrt{\frac{M}{N}}$, Apply Grover's operator (*G*) to states $|S\rangle$ for *k* times.

$$G^k|S\rangle = sin\,((2k+1)\theta)|\alpha\rangle + cos\,((2k+1)\theta)|\beta\rangle \quad (32)$$

When this appears, the target state will be explored with the probability of success *P*, formulated as follows:

$$p = sin^2((2k+1)\theta)) \quad (33)$$

Set $k = \frac{\pi}{4}\sqrt{MN}$, The Fig. 8 corresponds to the probability of success according to the proportion of target states in Grover's algorithm. To make things easier, let us set the proportion of the target as $\gamma = M/N$.

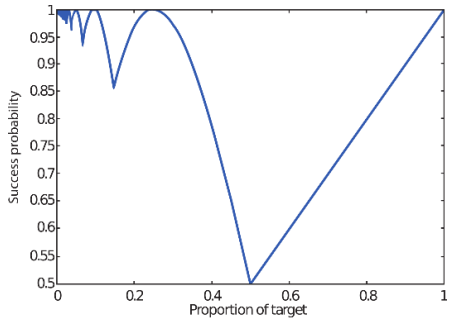


Fig. 8. The Success Probability of Grover's Algorithm.

To make things easier, let us set the proportion of the target as γ = M/N. Analyzing Fig. 8, we notice that the minimum probability that Grover's algorithm can reach is about 50% when γ = 0.5. Therefore, when 1/4 ≤ γ ≤ 1/2 the success probability of the proportion target declines rapidly. In return, when γ ≥ 1/2 the success probability of the proportion target gradually increases until it reaches 100% full accuracy when γ = 1.

C. Comparison of Grover's with Classical Algorithms

The practical implementation of Grover's search algorithm proved the efficiency in terms of its accuracy. After analyzing the different iterations, we found that the algorithm's effectiveness is influenced by the number of iterations. Moreover, applying the Grover algorithm iterations for a total number of *round* ($\frac{\pi}{4}\sqrt{N}$) times is the best choice to maximize the success probability of Grover's quantum search algorithm. Further, the quadratic reduction complexity of the quantum search Grover algorithm presents a major advantage over classical algorithms and exceeds any known classical algorithm of sub-exponential complexity. As shown in Fig. 9, Grover's quantum algorithm complexity time and classical counterpart algorithms.

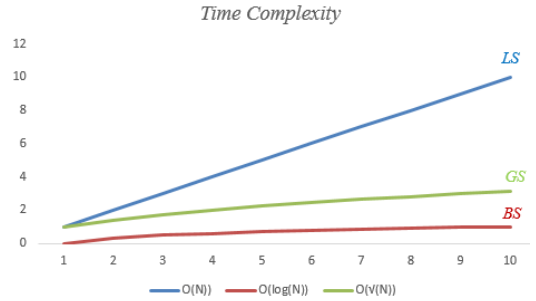


Fig. 9. The Average Number of Steps Needed to Find a Solution.

Therefore, a benchmark examination comparing the conventional search algorithms like sequential and interval search with the quantum Grover algorithm is still required. Table II shows a benchmark study between classical search algorithms and their counterparts, Grover's quantum search.

TABLE II. COMPARISON BETWEEN THREE DIFFERENT SEARCH ALGORITHMS

	Binary Search	***Linear Search***	***Grover Search***
Time complexity	*O(log(N))*	*O(N)*	*O(√(N))*
Database requirements	Database must be sorted	No requirements	No requirements
Algorithm type	Divide and conquer	Iterative	Iterative and parallel
Implementation	Medium	Easy	Hard

According to the comparison in Fig. 9, the binary search is the most sophisticated search, but it requires that the data be sorted, which is no longer possible with unstructured data. Linear search can override binary search if the targeted element exists at the beginning. However, being a search request for one or more elements in the heterogeneous database that contains data in different formats (structured, semi-structured, and unstructured), like the case of a data lake, Grover's algorithm remains the most efficient compared to the classical searching algorithms. Consequently, quantum algorithms are more prominent and highly recommended thanks to their quadratic acceleration, which is very fast compared to exponential acceleration, which corresponds to classical algorithms.

V. CONCLUSION

In this paper, an interesting algorithm is used to solve the search problem for unstructured datasets. We have investigated a clear procedure for making use of the potential of the quantum search Grover algorithm by proposing the design and implementation of the algorithm, including the prevalence effect of the number of iterations to decrease data processing time in unsorted databases. Based on this solution, our experimental results are very encouraging, and demonstrate the usefulness of Grover's algorithm to be applied efficiently to solve the search problem with high accuracy. Thus, from the benchmark search algorithms discussed in Section IV.C, we have retained that Grover's algorithm appears the best solution to the search problem in an unstructured data space. An important future perspective consists of moving to a higher dimension to solve the larger space search challenge with a large number of qubits.

REFERENCES

[1] Dabbèchi H, Nahla Z, Haytham E, Kais H. NoSQL Data Lake: A Big Data Source from Social Media. In: International Conference on Hybrid Intelligent Systems, pp. 93-102. Hybrid Intelligent Systems (2020).

[2] Oussous A, Benjelloun F, Lahcen A, Belfkih S. NoSQL databases for big data. In: International Journal of Big Data Intelligence. A (2017). https://doi.org/10.1504/IJBDI.2017.085537.

[3] Dabbèchi H, Nahla Z, Haytham E, Kais H. Social Media Data Integration: From Data Lake to NoSQL Data Warehouse. In: International Conference on Intelligent Systems Design and Applications. A (2021). https://doi.org/10.1007/978-3-030-71187-0.64.

[4] Ashley M. Quantum algorithms: An overview. In: npj Quantum Information. A (2016). https://doi.org/10.1038/npjqi.2015.23.

[5] Qiskit. https://qiskit.org/. Accessed 01 December (2021).

[6] Huai-Chun C, Hsiu-Chuan H. Digital quantum simulation of dynamical topological invariants on near-term quantum computers. In: Journal of Quantum Information Processing. A (2022).

[7] Aimeur E, Gilles B, Sèbastien G. Machine Learning in a Quantum World. In: Conference of the Canadian Society for Computational Studies of Intelligence, pp. 431-442. (2006).

[8] Songfeng L, Braunstein L. Quantum decision tree classifier. Quantum Information Processing. A (2013). https://doi.org/10.1007/s11128-013-0687-5.

[9] Ouedrhiri O, Banouar O, El hadaj S, Raghay S. Intelligent recommender system based on quantum clustering and matrix completion. In: Concurrency and Computation Practice and Experience; 2022.

[10] Dixon, J (CTO of Pentaho). Hadoop, and Data Lakes. In: Dixons Blogs. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/. Accessed 21 Sep 2021.

[11] Sawadogo P, Darmont J. On data lake architectures and metadata management. In: Journal of Intelligent Information Systems. A (2021).

[12] Inmon B. Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump. Bill Inmon - Google Livres. (2016).

[13] Khine P, Wang Z. Data lake: a new ideology in big data era. In: ITM Web of Conferences. A (2018).

[14] Cherradi M, El Haddadi A, Routaib H. Data Lake Management Based on DLDS Approach. In: Networking, Intelligent Systems and Security, pp. 679-690. (2022).

[15] Hellerstein et al. Ground: A Data Context Service. CIDR (2017).

[16] Hegazi O M, Saini K D, Zia K. Moving from Heterogeneous Data Sources to Big Data: Interoperability and Integration Issues. In: International Journal of Advanced Computer Science and Applications(IJACSA), Volume 9 Issue 10, 2018.

[17] Cherradi M, EL HADDADI A. Data Lakes: A Survey Paper. In book: Innovations in Smart Cities Applications Volume 5. January (2022).

[18] Mavroeidis V, Vishi K, Zych D M, Jøsang A. The Impact of Quantum Computing on Present Cryptography. In: International Journal of Advanced Computer Science and Applications(IJACSA), Volume 9 Issue 3, 2018.

[19] Anton F. Quantum optics with artificial atoms: Thesis for: PhD. (2014).

[20] Brian R, Classical emulation of a quantum computer. In: International Journal of Quantum Information. Vol. 14, No. 04, 1640004 (2016).

[21] Lov K. A fast quantum mechanical algorithm for database search. Computer Science, Physics. A (1996). https://doi.org/10.1145/237814.237866.

[22] Peter W. Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer. Quantum Physics. A (1996). https://doi.org/10.1137/S0097539795293172.

[23] Shor P. Algorithms for quantum computation: discrete logarithms and factoring. In: Proceedings 35th Annual Symposium on Foundations of Computer Science. (1994).

[24] Mandviwalla A, Keita O, Bo J. Implementing Grover's Algorithm on the IBM Quantum Computers. In: IEEE International Conference on Big Data, pp. 701-710. Springer (2018).

[25] Kyongyob M. Dirac Bra-ket Notation for Interpreting Regional Distribution of Pulmonary Ventilation-Perfusion. International Journal of Teaching and Case Studies. (2015).

[26] Tin T. The Efficiency of the Quantum Search : How effective is Grover's algorithm (quantum search) opposed to classical computer searching algorithms in terms of time complexity. A (2020). https://doi.org/10.13140/RG.2.2.18744.57604.

[27] Zubairy M. Quantum Superposition and Entanglement. Quantum Mechanics for Beginners. A (2020). https://doi.org/10.1093/oso/9780198854227.003.0010.

[28] Ryan L. Overview and Comparison of Gate Level Quantum Software Platforms. Computer Science, Mathematics, Physics. (2018).

[29] Ian M. IBM ups the stakes for Quantum Computing. https://www.enterprisetimes.co.uk/2017/11/13/ibmups-stakes-quantum-computing/. Accessed 07 October 2021.

[30] IBM, Corporation: IBM Quantum Experience. https://quantum-computing.ibm.com/. Accessed 12 August 2021.

[31] IBM, Corporation.: IBM Quantum Documentation. https://quantum-computing.ibm.com/docs/. Accessed 3 September. 2021.

[32] Marco M, Enrico P.A continuous rosenblatt quantum perceptron . In: International Journal of Quantum InformationVol. 19, No. 04, 2140002 (2021).

[33] Komal S. An Indexed Sequential Search and its Comparative Analysis with basic Searching Techniques. IJEAST. A (2020). https://www.ijeast.com/papers/559-564,Tesma504,IJEAST.pdf.

[34] Kostakis O, Gionis A.Subsequence Search in Event-Interval Sequences. ACM. SIGIR. A (2015). https://10.1145/2766462.2767778.

[35] Arkadiusz L. Rafa l R. Quantum Digital Signatures for Unconditional Safe Authenticity Protection of Medical Documentation. HIGHER SCHOOL'S PULSE. A (2015). https://doi.org/10.5604/2081-2021.1191752.

[36] You Q, Franco N. Superconducting Circuits and Quantum Information. Quantum Physics. A (2005). https://10.1063/1.2155757.

[37] Arkadiusz L, Laurentiu N. The Research of Grover's Quantum Search Algorithm with Use of Quantum Circuits QX2 and QX4: Part I. In: Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology – ISAT, pp 146-155. ISAT (2019).

[38] Panjin K, Daewan H, Kyung C. Time–space complexity of quantum search algorithms in symmetric cryptanalysis: applying to AES and SHA-2. In: Journal of Quantum Information Processing. A (2018).

[39] Gernot S. Quantum algorithm for optical template recognition with noise filtering. Physical review A, Atomic, molecular, and optical physics. A (2006). https://doi.org/10.1103/PHYSREVA.74.012303.

[40] Ulyanov S, et al. Modelling of Grover's quantum search algorithms: implementations of Simple quantum simulators on classical computers. In: Computer Science Journal. A (2020).

[41] Riccardo F. Quantum Amplitude Amplification Algorithm: An Explanation of Availability Bias. In: Proceedings of the 3rd International Symposium on Quantum Interaction, pp. 84-96. (2009).

[42] Akanksha S, Arko C. Grover's Algorithm. Architecture Design and Implementation of the Quantum search algorithm. A (2018). https://doi.org/10.13140/RG.2.2.30860.95366.

Inclusive Study of Fake News Detection for COVID-19 with New Dataset using Supervised Learning Algorithms

Emad K. Qalaja[1], Qasem Abu Al-Haija[2]*, Afaf Tareef[3], Mohammad M. Al-Nabhan[4]
Department of Computer Science, Mutah University, Karak, Jordan[1]
Department of Computer Science/Cyber Security, Princess Sumaya University for Technology (PSUT), Amman, Jordan[2, 4]
Department of Information Technology, Mutah University, Karak, Jordan[3]

***Abstract*—Covid-19 imposes many bans and restrictions on news, individuals and teams, and thus social networks have become one of the most used platforms for sharing and destroying news, which can be either fake or true. Therefore, detecting fake news has become imperative and thus has drawn the attention of researchers to develop approaches for understanding and classifying news content. The focus was on the Twitter platform because it is one of the most used platforms for sharing and disseminating information among many organizations, personalities, news agencies, and satellite stations. In this research, we attempt to improve the detection process of fake news by employing supervised machine learning techniques on our newly developed dataset. Specifically, the proposed system categorizes fake news related to COVID-19 extracted from the Twitter platform using four machine learning-based models, including decision tree (DT), Naïve Bayes (NB), artificial neural network (ANN), and k-nearest neighbors (KNN) classifiers. Besides, the developed detection models were evaluated on our new dataset, which we extracted from Twitter in a real-time process using standard evaluation metrics such as detection accuracy (ACC), F1-score (FSC), the under the curve (AUC), and Matthew's correlation coefficient (MCC). In the first set of experiments which employ the full dataset (i.e., 14,000 tweets), our experimental evaluation reported that DT based detection model had achieved the highest detection performance scoring 99.0%, 96.0%, 98.0%, and 90.0% in ACC, FSC, AUC, and MCC, respectively. The second set of experiments employs the small dataset (i.e., 700 tweets); our experimental evaluation reported that DT based detection model had achieved the highest detection performance scoring 89.5%, 89.5%, 93.0%, and 80.0% in ACC, FSC, AUC, and MCC, respectively. The results obtained for all experiments have been generated for the best-selected features.**

Keywords*—Machine learning; fake news; twitter; covid-19; correlation coefficient

I. Introduction

Over the years, many researchers have tried to identify fake news spreading on social media. Fake news is a source of spam capable of influencing perception, knowledge, and measuring methods [1]. Fake news has the potential to reach individuals through social media, cause damage to the economy and manipulate political outcomes. Fake news can be described as misinformation directed to deceive people [2]. In recent years, fake news has been shared on various social media. Generate a health concern to obtain advertising revenue for financial or political gain. When a particular news story is published, supporters of the news tend to share complete information without any falsification. However, those whose opinions do not correspond to the mentioned information. They resort to sharing the same information with some modifications of their own. As a result, the distinction between real and fake news has gained the attention of organizations such as Facebook, Google, and Twitter. Many researchers are making sustained efforts to combat the spread of fake news. Understanding the language in news stories is difficult because different people understand language differently. That is why the same news can be considered real or fake by a different group of people. The spread of fake news on these platforms leads to a loss of credibility and financial loss.

In 2019, a new virus called Covid-19 was reported in Wuhan, China, and the Covid virus has spread to various other parts of the world and has killed many people. At first, it was claimed that it was transmitted from animals to humans. Research and various experiments to find an effective treatment for covid-19 has become a very urgent need. Covid-19 has opened the door to spreading false news on various social media platforms such as Twitter, Facebook, and Instagram, which has misled many users worldwide. Misleading information and news about the disease are shared on the Internet from various sources, some of which are not trusted. It is well known that spreading false news about Covid-19 on social media can contribute to stress and health anxiety and lead to serious consequences for society's awareness and reaction to vaccination against Covid-19, such as misinformation about false treatments, anti-vaccination propaganda, and theories of the plot.

With advancements in processing technology, machine learning models, and deep learning techniques, user intervention can be replaced by assigning pattern identification tasks to computers. On the other hand, very little research has been done on applying linguistic and deep learning techniques for accurate classification of fake news among research done; the accuracy achieved is so high. This paper discusses the classification of fake news related to Covid-19 using Machine Learning Algorithms (MLA) and will focus on the news spread on the social media platform Twitter. This is done by enhancing the process of detecting fake news using machine learning algorithms such as Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbors (KNN)

classifiers, and Artificial Neural Network (ANN) that can manage and distinguish real and fake news about Covid-19.

A. Problem Statement

Fake news in people's lives is a spam source that can affect people's general perception and knowledge [1]. Fake news can be described as a specific type of misinformation sent and directed to deceive people [2]. In recent years, a new term in the scientific arena is (electronic flies), especially with the massive development of digital media and communications and the spread of social media platforms and their direct and frightening impact on the behavior of individuals. Especially when directing society to a studied destination by publishing news to serve a specific issue, which leads to new terms such as digital propaganda, digital war, digital armies, and electronic terms. [3]. When a global news story is published, news sites and organizations race to share their coverage and stories on social media. Proponents tend to share information in its complete authenticity without change. However, those whose opinions do not correspond to the published information and may materially harm them, resort to sharing the same information with some modifications of their own, which leads to the existence of complete real news and fake news, which leads to confusion in people's understanding of the truth. Hence the distinction between real and fake news has received considerable research interest. And influential organizations, such as Facebook, Google, and Twitter, are making sustained efforts to combat the spread of fake stories. Since the start of the COVID-19 crisis, much false news has spread rapidly on social media about the disease, its symptoms, and the number of infections, as well as fake news about vaccines and their side effects. Detecting and distinguishing between real and fake news has posed a challenge to researchers regarding the accuracy of the results, the speed of obtaining them, and the stability of the technique used.

B. Research Contribution

The main contribution of this research is proposing a model to detect fake news on a Twitter platform using MLA and meta-date (attributes for a Twitter account). The model used MLA to build the behavior of members of evaluation panels and to resolve the multiplexing between their judgments. This study contributed the following:

- Collecting correct and fake tweets and corresponding metadata to create a dataset that will be publicly available for other researchers in the same field.
- Designing and developing an accurate model to detect fake Covid-19 news on Twitter using an MLA and Twitter's meta-data.

II. Literature Review

Due to the proliferation of large volumes of false content during the pandemic, the study around Covid-19-related misinformation became a popular area of research. Several methods were proposed to differentiate and verify the real and fake news for Covid-19 from different datasets and resources. Authors of [4] used deep learning algorithms in their study. The proposed model was based on the tweet's text and other tweet's features extracted online from Twitter, such as favorite count, retweet count, source, length, verified, the user URL, friend/followers count, statuses/followers count, and sentiment. The proposed method achieved an accuracy of 79% compared with SVM (72%) using Sheryl Mathias and Namrata Jagadeesh's dataset and the fake news data repository "FakeNewsNet." Recall reached 100% using RF, while the DT reached 94%. RF has 85% for the precision and 83% for the F1 score. In [5], the authors proposed a system for fake detection news consisting of two main categories: MLA and DNL. He used the FakeNewsNet dataset containing news content, social context and spatiotemporal and disasters, PolitiFact, and gossip police information to identify fake news on social media. The performance measure results were as follows: LSTM (Two layers) regarding disasters dataset (accuracy 98.6%, precision 98.55%, recall 98.6%, F1-score 98.5%). The Modified LSTM (one layer) obtained the best testing results: regarding the disasters dataset (accuracy 86.74%, precision 86.98%, recall 86.74%, F1-score 86.6%).

Regarding the PolitiFact dataset, the best testing results are obtained by the modified LSTM (two layers) (accuracy 83.93%, precision 86.66%, recall 83.93%, F1-score 83.31%). Regarding the gossip police dataset, and finally the Modified LSTM (one layer) regarding gossip police dataset (accuracy 83.82%, precision 84.85%, recall 83.82%, F1-score 83.7%). In [6], they applied several NNs, LSTMs, ensemble methods, and attention mechanisms to detect fake news on Twitter and other media platforms. Their models for fake news classification are based on the sentiment analysis of users in social media. They used the architectures to detect patterns in their data, where patterns can be anything such as unusual capitalization, random exclamations, question marks, etc. Various datasets were also used for evaluation, like the PolitiFact dataset, FakeNewsNet dataset, and twitter's advanced search functionality. The results showed that the LSTM achieved the highest accuracy: 88.78%. The detection performance was 73.29% in the CNN, 80.62% in the LSTM, 83.81% in the bidirectional LSTM, 88.78% in the CNN + Bidirectional LSTM, and 57.58% in logistic regression.

In [2], they proposed a Fake news tracker to identify false news and prevent propagation. Deep learning models were used to classify the encoder site consistent with deep LSTMs with two layers and 100 cells. The obtained accuracy on the PolitiFact dataset was 63.3% and 74.2% on the Buzzfeed dataset. In [7], the researchers used two MLAs: SVM and RF. They achieved the best result on SVM: precision of 50%, recall at 30%, and F1 score at 60%. On the other hand, RF achieved a precision of 88%, recall of 89%, and F1 score of 89%. In [8], the contributors used a supervised learning classification to train and test the manually and automatically annotated datasets to ensure annotation quality. The proposed method includes six different ML algorithms, four different features with each algorithm, and three pre-processing techniques. This method achieved: an 87.8% F1-score classification result with the manually annotated corpus, the automatically annotated corpus F1-score of 93.3%, and the highest precision value was obtained using the n-gram TF-IDF feature with the LR classifier (87.8%), finally LR classifier (93.4%) on manually and automatically annotated corpora.

On the other hand, in [9], the authors used six machine learning algorithms: NB, KNN, RF, C4.5, BN, and SVM. The

train is based on a 10-fold cross-validation model. This study created a dataset of tweets collected using Twitter's streaming API spanning three months. The average accuracy of the cross-validation model for C4.5 was up to 98%, followed by RF, which had an average accuracy of 97.4%. The C4.5 outperformed all the other models. The Naive Bayes algorithm had the worst performance, with an average accuracy of 85.5%. In [10], they used three MLAs: NB, LR, and SVM, with two features: word embedding and word frequency approach. At the practical level, they collected one million Arabic tweets from the Twitter streaming API related to Covid-19. This study found that ML classifiers can correctly identify fake news-related tweets with an accuracy of 84%. In [11], the authors found that J48 has performed the best for the BuzzFeed Political News dataset with an accuracy of 0.655, while Classification via Clustering (CVC) has the worst accuracy of 0.501. For the Random Political News dataset, Sequential Minimal Optimization (SMO) algorithm has the highest value among the twenty-three algorithms, with an accuracy of 0.680.

In the same context, [12] discussed methods for detecting fake news using different sets of features extracted from the news text. One of the used feature sets was stylometric features, including the presence of uppercase letters and quoted content. Such features can be significant for detecting fake news and highlighting the importance of the writing style of news. Also, the write prints feature set extracted contains the content-specific, structural, linguistic, and syntax-based features. The model achieved an accuracy of 86% for stylometric features with a gradient boosting classifier. In [13], the authors used Bag-of-Words and TF-IDF, syntactic and semantic-based using Word2Vec and FastText. This method used two datasets for testing, and the results showed that the SVM model using TF-IDF obtained the best F1-Score value in both testing data. The model obtained an F1-Score of 92.21% in Testing Data 1 and 93.33% in Testing Data 2. In [14], the researchers tried to detect fake news using deep learning techniques such as LSTM, CNN, and BERT. The obtained accuracy results were LSTM 91%, CNN 93%, and BERT 98%. While in [15], they used four MLA classifiers: LR, SVM, DT, and Gradient Boost, to perform a binary classification to detect fake news and benchmark the annotated dataset. The proposed method curated and released a manually annotated dataset of 10,700 social media posts and articles concerning Covid-19 news, and it achieved the best performance of 93.32% F1-score with SVM.

Other noticeable models were found in [16-20]. In [16], machine learning was utilized to detect fake news published through social media such as Twitter and Facebook. The used ML algorithms were NB, SVM, BERT fine-tuning, and SBERT. The experiments found that SVM achieved the best results with F1 Validation of 93.28, compared to 90.62 using NB, 80.88 using BERT, and 78.18 using the SBERT technique. In [17], they proposed a detection method to distinguish and verify the fake news for Covid-19. This method achieved accuracy with the DT classifier at 92.07%, and the RF classifier accuracy achieved 94.49%. They proposed a model to classify news within different categories using SVM and TF-IDF. The classification precisions were 97.84% and 94.93% for BBC and 20 Newsgroup datasets. Also, in [19], the authors detect fake news in Covid-19 using a linear SVM, RF, LR, NB, and MLP. The evaluation was conducted using a large dataset containing 10,700 manually annotated social media posts and articles. The results showed that SVM achieved the best performance with 95.70 accuracies compared to others. SVM 95.7%, RF 90.79%, LR 95.42%, NB 93.32%, MLP 93.60%. In [20], they utilized an n-gram classifier to detect fake news. The TF-IDF feature extraction method estimated RF, DT, and SVM. This method achieved an accuracy of 0.73 for SVM and 0.78 for passive-aggressive.

Moreover, in [21], they used an n-gram classifier to detect fake news. SVM was estimated with the TF-IDF feature extraction method. The accuracy achieved 0.92. in [22], the authors used ten MLAs with seven feature extraction techniques to detect fake or real news. They tested their proposed classifier on 3,047,255 tweets concerning Covid-19. The best performance measures they achieved in NN, DT, and LR classifiers, were 99.7%, 99.9%, and 99.8%, respectively. In [23], they utilized two fundamental ML classification techniques within the meaning of text analytics. They identified common sentiments attached to the pandemic using the Coronavirus (COVID-19) Tweet and R analytical software. As Covid-19 approached the top level in the USA United States used clear textual analytics carried through needed text data visualization. The proposed method accuracy achieved 91% for long Tweets, including the Naïve, and an accuracy of 74% with a shorter tweets. While in [24], the study attempts to realize the rationale behind people's use of certain media, which was extended by an "altruism" motivation. The data were analyzed with Partial Least Squares (PLS) to determine the effects of six variables on the outcome of fake news. The researchers used Nigerian citizens as study samples, and the dataset contained 385 samples used in the experiments. The study showed that altruism is the most significant predictor of fake news sharing without using machine learning techniques. Furthermore, in [25], the researchers collected 2.7M posted by over 690k unique users. They noted that 18.66% of the tweets were posted by verified users (who constitute only 0.81% of the unique users). They collected 748k Arabic Language Tweets in addition to propagation networks of a subset of 65k Tweets to enable the research related to natural language processing, information retrieval, and social network analysis. This method used Twitter search API to retrieve the data daily between (January 27, 2020–March 31, 2020). The study did not use any MLA on the study and did not supply any results related to evaluation results. In addition, the collaborators of [26] collected a dataset containing 4072 news articles from Webhose.io regarding fake news about Covid-19. This method used linguistic features and conducted experiments with baseline classifiers, LSTM, and dense layer. The proposed method’s accuracy was between 70% and 80%.

Eventually, by reviewing the literature, researchers focused on studying real/fake tweet detection using popular machine learning algorithms. Some researchers used DL and NLP to discover the nature of tweets. Researchers have achieved excellent results through machine learning algorithms (using natural languages). But natural languages differ in understanding from each other, so the published tweet/news may be true in a specific language and for people who know

the details of the language, while the tweet/news is misleading for people who do not understand the language in which the news/tweet is published. On the other hand, the results that can be obtained using (NLP) can be obtained similar results if using machine learning algorithms with (metadata) provided by Twitter API. Note that the authors of [4] have used common MLA and DL algorithms and reached excellent results using common machine learning algorithms and (metadata). From our point of view, I think using common machine learning algorithms is sufficient if their results are excellent compared to the results reached by researchers when using DL algorithms. The author will use them during this study and compare them with the results of [4]. In this study, a proposed model will be presented that uses machine learning algorithms and Twitter metadata to improve fake news detection and real news by identifying features that affect the accuracy of results.

III. Materials and Method

This section presents the research methodology and the steps that were followed to achieve the goal and objectives of this research. The proposed approach is decomposed into data collection, feature selection, machine learning implementation (classification), and metric evaluation. Fig. 1 summarizes the steps of the proposed system.

A. Dataset Collection

This study will collect a data set using the Twitter API. To use Twitter's metadata, the metadata will be used as features of the dataset. It is one of the most important contributions of this research study, as the data set available on the different platforms provides a data set consisting of the tweet and the status of the tweet only (0/1, true/false) and does not provide the metadata that we need for the study. To implement and train the proposed model. We need a labeled data set, which enables the data set (tweets) to be sent to medical bodies specializing in Covid-19 to determine the type of tweet that is healthy/false. The dataset usually contains various forms of text, numbers, and language combinations, as well as some retransmission hashtags or tags; our dataset is extracted from the social media giant (Twitter) and used to detect fake news from real news after selecting key features from the data descriptive and humane evaluation of the data set by staff with medical backgrounds to determine appropriate features subsequently.

B. Feature Selection

Feature selection is a good way to infer features with a strong and effective effect, which improves accuracy results. The algorithm's time is not wasted on non-valued features. Many feature selection methods are available in the literature based on the abundance of data with hundreds of variables leading to high dimensional data. Feature selection methods provide a way to improve prediction performance, reduce computation time, and better understand a data set in machine learning or pattern recognition applications [27]. We can define a feature as an individual measurable characteristic of the experimental process. Through a combination of features, any machine learning algorithm can perform classification. Also, feature selection aims to select a small subset of relevant features from the feature pool obtained by removing inappropriate, redundant, or worthless/annoying features. [28]. There are common search strategies to select features, such as Information gain using a univariate information filter class applicable to classification [29], Minimum redundancy and maximum relevance: using a multivariate information filter class applicable to classification [30], and Correlation: using univariate information filter class applicable to regression [31], Correlation-based feature selection (CFS): using multivariate information filter class applicable to classification, regression [31], Fisher score: using univariate information filter class applicable to classification [28], and Spectral feature selection (SPEC) and Laplacian Score (LS): using univariate information filter class applicable to classification [30]. Based on our study and experiments and using Correlation-based feature selection (CFS), we noticed that some features do not affect the accuracy of the results even if they are excluded. For example, the gender and nationality of a news writer do not affect human opinion when checking the authenticity of real news from fake news. In this study, the feature selection is based-on correlation and ranking, as will be explained with an example in the next section. We examined each feature with the target "class," recorded all results, and compared results to each other to select the best features and then used these features on our proposed model. Our dataset had thirty-five features before medical panel validation (which will be discussed later).

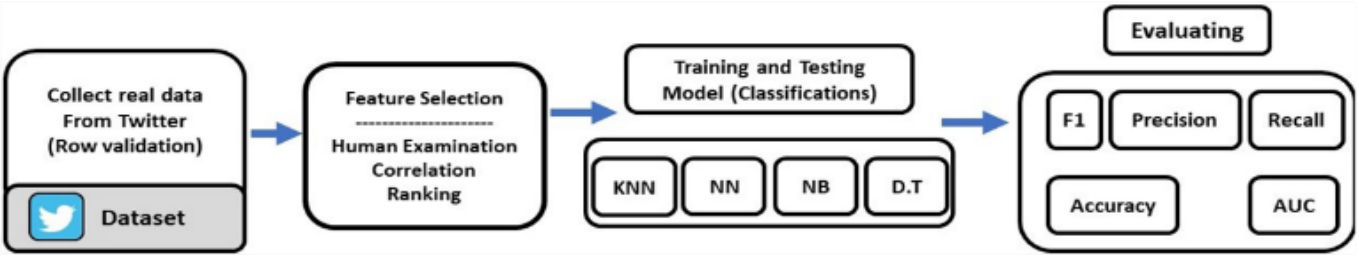


Fig. 1. The Proposed Model Architecture.

C. Machine Learning Implementation (Classification)

This study will collect a data set using the Twitter API. To use Twitter's metadata, the metadata will be used as features of the dataset. It is one of the most important contributions of this research study, as the data set available on the different platforms provides a data set consisting of the tweet and the status of the tweet only (0/1, true/false) and does not provide the metadata that we need for the study.

1) DT parameters:

- The minimum number of cases in papers where the algorithm will not create a division less than this limit, which would put less than the specified number of training examples in any branches.
- Split subset, Sub-division, where the algorithm is divided by a given number of instances.
- Tree depth limits the classification tree's depth to the specified node number.
- The majority (%): the algorithm depends on the division of the contract after reaching the specific majority threshold.
- Induce and build a binary tree (split into two child nodes).

2) KNN parameters:

- Distance Metric calculates the distance of 1 test observation from all other observations of the training dataset and then finds K nearest neighbors. To calculate the distance, we can use the following not exclusively: "Manhattan," which is the sum of all attributes' absolute differences, of all attributes, or "Mahalanobis," which is the distance between point and distribution. Or "Euclidean," which is the distance between two points, or "Chebyshev," which is the greatest of absolute differences between attributes.
- Weight: has two types: "Distance" is the closest neighbors of a query point have a greater influence than the neighbors further away, and "Uniform" is all points in each neighborhood are weighted equally. [32].

3) ANN parameters:

- Neurons are defined as the element that represents the number of neurons in the hidden layer. e.g., a neural network with three layers can be defined as 2, 3, 2.
- Activation is divided into "Logistic," the logistic sigmoid function. "Identity" is the no-op activation useful to implement linear bottleneck. "ReLu" is the rectified linear unit function. "Tanh" is the hyperbolic tan function.
- Regularization parameter alpha default value 0.0001.
- The solver for weight optimization contained "SGD "stochastic gradient descent."L-BFGS-B" is an optimizer in the family of quasi-Newton methods. "Adam" is a stochastic gradient-based optimizer that works relatively well with thousands of training samples or more in terms of training time and validation score. However, "L-BFGS-B" can converge faster and perform better.
- A Maximal number of iterations is 200 [32].

D. Evaluating Metrics

It is now well known that error rate is not an appropriate evaluation criterion when there are unequal costs. This paper uses F-measure and AUC (Area under the ROC Curve) as performance evaluation measures.

1) F1-measure is the mean of precision and recall. This takes the contribution of both, so the higher the score, the better, as shown in equation 2. The F1-measure is calculated by multiplying (Precision and Recall by 2) value divided by the total of precision and Recall.

$$\mathrm{F-measure} = (2 \times \mathrm{Prec} \times \mathrm{Rec})/(\mathrm{Prec} + \mathrm{Rec}) \quad (1)$$

2) AUC has proved to be a reliable performance measure for imbalanced and cost-sensitive problems. Given a binary classification problem, a ROC curve depicts the performance of a method using the (FP, TP) pairs. FP is the false positive of the classifier, and TP is the true positive. AUC is the area below the curve [33]. The calculation for FP and TP is shown in equations three and 4.

$$False\ Positive\ (fp) = FP/(FP + TN) \quad (2)$$

$$TruePositive\ (tp) = TP/(TP + FN) \quad (3)$$

3) The confusion matrix is a table that illustrates and displays a performance rating model on a data set whose true values are already known. It is the best way to understand the behavior of the technique and algorithm used to show the statistics and the relationship between the expected results. As shown in Table I.

TABLE I. CONFUSION MATRIX

	Predicted Positive	**Predicted Negative**
Actual Positive	(True Positives)	(False Negatives)
Actual Negative	(False Positives)	(True Negatives)

4) Precision is an evaluation metric measuring the percentage of positive cases out of the expected positive cases. Equation 5 shows how to calculate Precision. [34].

$$Precision = TP/(TP + FP) \quad (4)$$

5) The recall is the part of the relevant documents that have been successfully retrieved. The recall is calculated as shown in equation 6 [34].

$$Recall = TP/(TP + FN) \quad (5)$$

6) Accuracy is the most used metric to judge models, which is calculated by summation of true positive values and true negative values divided by summation of true positive and false positive and true negative and false negative,

according to equation 7, Where the values of TP, TN, FP, and FN were taken from the confusion matrix.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (6)$$

7) Matthew's correlation coefficient (MCC) is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between "-1", "0" and "1" as the following explanation: Correlation of "0" shows no linear relationship between the movements of the two variables, the Correlation number is greater than "1" or less than "-1" means that there was an error in the correlation measurement, and Correlation of "-1" shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. [35]. Matthew's scale is associated with the F1 scale. As the F1 scale rises, the higher the Matthew scale rises.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (7)$$

IV. Results and Discussion

This section presents five stages of the experimental setup: experimental environment, dataset collection, cleaning dataset and pre-processing setup, dataset evaluation, result, and analysis (selection features and finally presents the accuracy results), recommendation, and future planning work.

A. Experimental Environment

As an intelligent adaptive approach, several phases describe the relationship between phases, where the outputs from a specific phase can be considered inputs for the following phase. In addition, moving to the next phase should ensure that the previous phase is completed. And to ensure that it will reach the best result, all ML algorithms and computational techniques were performed on a personal computer with an Intel(R) Core (TM) i5-2410M CPU @ 2.30GHz, with 8 GB of RAM and 512 GB SSD hard disk.

B. Dataset Collection

The dataset was obtained from the social media platform (Twitter) Orange App ver. 3.30.2 Using the text extraction extension, this program helps extract tweets easily by obtaining permission from Twitter to use its tweets in scientific research. Operations (correlation/ranking, classification, and all evaluation results) were performed through Python version 3.8. Orange is the software of a Python-based component. Visual programming software for data mining, machine learning, and data analysis. Data is presented visually, and App allows classification and clustering. Table II shows useful details for getting tweets using Orange App and Twitter Add-on; as we explained earlier, Orange was used as a program through which tweets are fetched using the secret key granted by Twitter.

We contacted Twitter to get API Key and developer account to search and collect real tweets from the original resources. This method also provided access to Twitter attributes used as features in this study. Table III displays a set of meta-data (Attributes) from the Twitter App using API and python command to get a user objects directory, known as Twitter's (metadata). For example (followers count, retweet count, likes, time, author- Verified, username, ID), this study focused on testing metadata as features to find out which (features) achieved the best Accuracy results F- measures, recall, Precision, MCC.

TABLE II. Twitter API Variables

Variables	Description
Twitter API Key	API key, secret, token, and Bearer token are simple encrypted string that identifies an application without any principal
API Key Secret	
Access Token	
Bearer Token	
Query Word list	Hashtag you are searching for, e.g., COVID-19
Languages	Searching by language (English, Arabic…etc.)
Max tweets	The maximum tweets count
Search type:	default "content, Target search type

After that, we collected 14,000 tweets, as shown in Table IV, and a sample of 675 tweets was taken for the study. The dataset will be available to researchers for research purposes and research studies. The dataset has been cleaned and prepared to get features used in our proposed model and applied machine learning algorithms to it. To exclude useless features such as (date, id_str, id, entities, user, longitude, latitude, user_truncated, place, user_producted, user_description), which have frequent data values, and delete unwanted row heads from the dataset, an ML test must be implemented to dataset after cleaning.

TABLE III. List of Attributes

Attribute	Description
user_statuses_count	The number of tweets and retweets made by the user
user_listed_count	represents the number of public lists this user is a member of (registered as a member of the list/group)
user_followers_count	The number of followers the account currently has
user_protected (if true)	Indicates that user has chosen to protect his Tweets
user_description	represents the description of the current user account
Id	Integer-unique number of this user (the current user)
Number of Likes	The number of likes that are recorded on a specific tweet for the current user
created_at	The date on which the current user account was created on the social networking platform (Twitter)
user_location	The location that the user (the account holder) specified for this account
Tweet	The text of Tweet posted on Twitter for current user
source_url	URLs included by user in the text published Tweet
Lang	The language the user registered as the mother tongue in the account upon creation
in_reply_to_user_id_str	If this field contains a string of the original Tweet owner ID, the represented Tweet will be a reply
in_reply_to_status_id_str	If this field contains a string of the original Tweet owner ID, the represented Tweet will be a reply

Attribute	Description
in_reply_to_screen_name	If this field contains the screen name of the original Tweet author., the Tweet will be a reply
Entities	The set of features extracted from the Tweet body (user_mentions, hashtags, media, symbols)
user_favourites_count	The number of times of users liked a particular tweet
user_friends_count	The number of users who follow a particular account
user_verified (if true)	Indicate that the user has a verified account
number of retweets	The count of a user has retweeted a specific Tweet
URL	URL contained in the text of the published Tweet.
truncated	If the retweet exceeds the text length limit of the original Tweet, the value of this field is valid and is expressed as a Boolean value.
Source	it is a utility for posting a Tweet and is expressed as a string formatted in hypertext language (HTML). It indicates the source from which the tweet was launched (a website or application from a smartphone, whether Android or iPhone)
Id_str	each Twitter user is unique and has an identifier string, represented by this field, and is considered a property of the user (tweeter)
User	it represents an object describing the user who posted the tweet. It has several attributes: (screen_name, location, screen_name)
geo_enabled	this field indicates that the user has allowed the ability to geotag their Tweets, if true.
retweet_count	this field indicates the number of times a Tweet was retweeted by any user who posted a Tweet
is_quote_status	this field indicates whether the Tweet was quoted
in_reply_to_user_id	Contains a valid representation of the original Tweet owner ID, so the tweet represented was a reply.
in_reply_to_status_id	If contains a valid representation of the original Tweet owner ID, the tweet represented was a reply.
favorite_count	The number of tweets liked by the user over the life of the account
display_text_range	The length of the text of the tweet, bearing in mind that the Twitter platform imposes on the user a specific number of characters for a single tweet, and it cannot be exceeded within the same tweet
default_profile	It represents a Boolean value and indicates that the user has not changed the background of his user profile; if the value of the field is true, and if the value is false, then the user has changed the background of the image
Has_background_image	If the field value is true, the user wants to use the uploaded background image.

TABLE IV. THE DATASET INFORMATION

	Column	Non-Null Count
1	display_text_range	14000 non-null
2	entities	14000 non-null
3	favorite_count	14000 non-null
4	in_reply_to_screen_name	13898 non-null
5	in_reply_to_status_id	14000 non-null
6	in_reply_to_status_id_str	14000 non-null
7	in_reply_to_user_id	3413 non-null
8	in_reply_to_user_id_str	13996 non-null
9	is_quote_status	12791 non-null
10	lang	14000 non-null
11	retweet_count	14000 non-null
12	Number of Retweets	14000 non-null
13	place	14000 non-null
14	Number of Likes	14000 non-null
15	source_url	14000 non-null
16	geo_enabled	13898 non-null
17	Tweet	11351 non-null
18	User	14000 non-null
19	user_location	14000 non-null
20	Id_str	14000 non-null
21	created_at	14000 non-null
22	source	14000 non-null
23	truncated	14000 non-null
24	id	14000 non-null
25	URL	14000 non-null
26	user_description	14000 non-null
27	user_protected	14000 non-null
28	user_verified	14000 non-null
29	user_followers_count	14000 non-null
30	user_friends_count	14000 non-null
31	user_listed_count	14000 non-null
32	user_favourites_count	14000 non-null
33	user_statuses_count	14000 non-null
34	profile_use_background_image	14000 non-null
35	user_default_profile	9677 non-null

C. Cleaning Dataset and Pre-processing Setup

The data cleaning process started by removing the empty rows and, incomprehensible symbols, useless attributes (date, time, language, latitude, longitude, in reply to, and location). Also, change feature values from Boolean to numeric (true/false → to 1/0).

D. Dataset Evaluation

After collecting and revising the dataset and having 675 tweets in finalizing step; the following steps were followed:

Some personal tweets were excluded (for example … Pfizer's second graft dose was taken, and I am on my way to take a booster dose of grafts...) to reach 542 tweets. For the next step. The medical panel validated the tweet content as fake/real news by humanly evaluating those with medical backgrounds by sending the dataset, which contained 542 tweets, to the medical panel. A data set consisting of two fields has been sent (Tweet, True/False rating) to get the medical opinion and evaluation tweets and select it as "Class" for the dataset and other attributes as features (Num-Likes, Num-Retweet, Author Followers Count, Author Listed Count, Author Favorites Count, Author Friends Count, Author Statuses Count and Author verified). The number 'True' will be selected for the real Tweet, and 'False' will be selected for the fake Tweet. The Tweet (content) will only be used to classify it, and then it will be excluded before it is entered into the classification algorithms; we exclude any feature that contains text such as the name of the Tweet author or the description of the Tweet author, and we convert the value (true/false) to "0" for the false tweet and "1" For the correct tweet, We combine it with the rest of the features (Num-Likes, Num-Retweet, Author Followers Count, Author Listed Count, Author Favorites Count, Author Friends Count, Author Statuses Count, and Author verified), to get the final dataset for the study.

E. Experimental Result Analysis

After collecting the datasets, author excluded the tweet from the dataset and applied the proposed model to test and train machine learning algorithms on it, as following steps:

Importing Dataset: In this step, import and read the data set, using Python commands and show data set information, and repeat false tweets and is represented at zero number as well as the correct tweets and are represented in one number and all of which are represented by Class. This means that the data set is almost balanced.

Correlation: In this step, Correlation was used through Python commands to get the most useful features in terms of interconnection between them. The features that most affect the result are identified by Dataset Correlation, one of the most important commands in the Python library because it identifies features that affect the accuracy of the results when machine learning algorithms are applied to the dataset. Remember that all the features have been checked with the class feature, as we will explain in the next step. Fig. 2 shows the important and best features that affect the accuracy of the results. Fig. 2 shows the important and best features were (Num-Likes, Num-Retweet, Author Followers Count, Author Listed Count, Profile use background images, user default profile) Where the result of the impact of features (Num-Likes, Num-Retweet, 0.98) and features (Author Followers Count, Author Listed Count,0.89) and features (Profile use background images, user default profile,0.59) in other words if we remove any one of correlated features the evaluation result will reduce.

Feature selection: We applied the proposed form in two stages: The First Phase is to find the best class. Due to the number of tweets, we obtained in a huge data set (14000 tweets) and where the medical team could not be done; Because the huge number of tweets began to test selected features of metadata to be our class, and all the features were checked, and the class we reached was the author verification, several likes, and re-tweet as a "class." The proposed form was applied to the data set with the ML CALSIFIERS (DT, KNN, ANN, NB, LR, RF, SVM). The results were unreasonable and can arrive in workbooks (RF and DT) at 9.9%, and the difference between the right and false tweets and which were "0" and "1" (as a class), was not balanced. It is clear through the data set results that we will have to balance the data again, leading to an increase in samples that do not exist or excluding samples affecting the accuracy we will receive. After that, randomized random sampling method and random sampling method were used to balance our data set, but we found that the random sample with excess factors had disadvantages (have increased the sample with non-realistic values), and this led to incorrect resolution for results, in addition, The random sampling method also defects (sample is deleted that may contain data affecting resolutions)—the results as shown in Table V.

The second phase: A sample of 670 tweets was taken after cleaning the data set and reading Tweets by removing unnecessary or unnecessary tweets. Where the revised final data set reached 543, the authors are keen to be balanced as much as possible, avoiding the problem of non-balanced data. The data set was sent to the specialized medical authorities to evaluate Tweets. Then, the authors applied the proposed form to the data set. The authors have done the following: It should be noted that each feature was examined with target "Class" one-by-one consequentially and recorded the results to reach the best results, and through this process, it was found that (Num-Likes, Num-Retweet, Author Followers Count, Author Listed Count) are the most influenced result in the accuracy and improved results and found that features: Author Favorites Count, Author Friends Count, Author Statuses Count and Author verified had reduced the accuracy results as Table VI and Table VII shown. However, the correlation and ranking method helped to find the best correlation features, which was achieved with our next step.

Training and Testing Dataset: The dataset was divided into 70% for training and 30% for testing using python; the Machine learning algorithms (DT, NB, KNN, NN) were used because it is the best classifier for (Binary dataset attributes) and also easy and fast classifiers. The parameters settings for Classifiers were as follows: (A) DT parameters:[minimum number of instances in leaves (10), the smallest subset(5), maximal tree depth (30), the majority reaches (95%)], (B) KNN parameters: [number of neighbors (5), metric (Euclidean), weight (Uniform)], and (C) ANN parameters: [Neurons(100,), activation (Relu), solver (Adam), regularization by default (0.0001), maximal number of iterations(200)]. Classifiers have been applied using Cross-validation by (3, 5, 10, 20) folds. The best result was achieved after applying machine learning algorithms to the dataset using Cross-validation with 20 folds as follows: Decision Tree (DT) and Naïve Bayes (NB) achieved the highest value of Accuracy in Evaluation Results it was 89.5%, K-Nearest Neighbors (KNN) achieved 88.9% value of Accuracy in Evaluation Results, the Neural Network (NN) has 82.1% in Evaluation Result of Accuracy as shown in Table VIII.

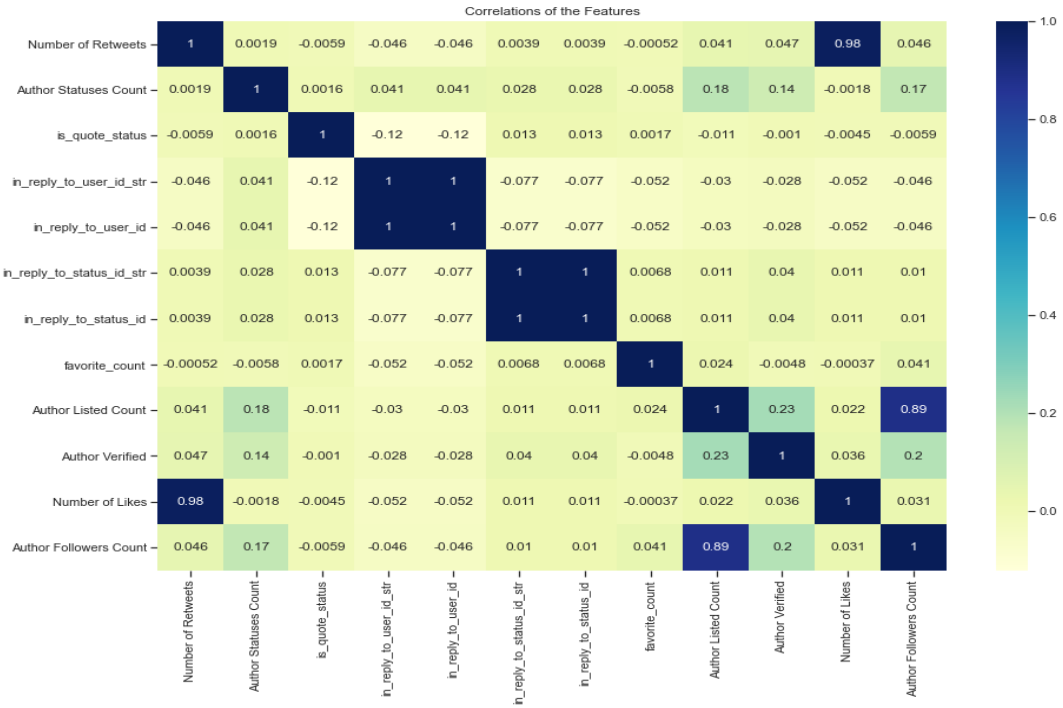


Fig. 2. Best Features Selected by Correlation.

TABLE V. RESULT OF 14000 TWEETS

Target	Model	TP	FP	FN	TN	CA	PR	RC	F1	AUC	MCC
A-V, Like, Retweet	KNN	1567	498	443	967	0.71	0.75	0.77	0.76	0.9	0.4
A-V, Like, Retweet	SVM	1982	1069	28	396	0.67	.0.64	0.98	0.78	0.8	0.3
A-V, Like, Retweet	DT	1942	61	68	1404	0.99	0.96	0.96	0.96	0.98	0.9
A-V, Like, Retweet	LR	1985	113	25	1352	0.9	0.94	0.98	0.96	0.98	0.9
A-V, Like, Retweet	RF	1989	61	21	1404	0.99	0.97	0.98	0.97	0.98	0.9
A-V, Like, Retweet	NB	1981	1145	29	320	0.65	0.63	0.98	0.77	0.71	0.3
A-V, Like, Retweet	ANN	931	71	1079	1394	0.72	0.92	0.46	0.61	0.78	0.4

TABLE VI. FEATURES REDUCED ACCURACY

Features	**Model**	**AUC**	**Acc**	**F1**	**Rcl**	**Prc**
Author Verified Author Friends Count Author Favorites Count Author Statuses Count **Target "**Class"	**ANN**	0.82	0.75	0.75	0.75	0.76
	KNN	0.79	0.74	0.73	0.74	0.74
	NB	0.80	0.72	0.72	0.72	0.72
	DT	0.67	0.63	0.63	0.63	0.63

TABLE VII. FEATURES ENHANCED ACCURACY

Features	**Model**	**AUC**	**Acc**	**F1**	**Rcl**	**Prc**
Author Followers Count Author Listed Count Number of Likes Number of Retweets **Target "**Class"	**ANN**	0.92	0.89	0.89	0.89	0.89
	KNN	0.96	0.89	0.89	0.89	0.89
	NB	0.92	0.88	0.88	0.88	0.88
	DT	0.89	0.82	0.82	0.82	0.83

TABLE VIII. EVALUATION RESULT FOR CLASSIFIERS

Model	**AUC**	**Acc**	**F1**	**Rcl**	**Prc**
DT	0.928	0.895	0.895	0.895	0.897
NB	0.960	0.895	0.895	0.898	0.899
KNN	0.923	0.889	0.889	0.889	0.889
ANN	0.894	0.821	0.821	0.821	0.839

On the other hand, this proposed model is a classification based on meta-data because of the advantages of using meta-data (it can represent directly and easily). The results of testing classification algorithms plots for the AUC-ROC curve, as shown in Fig. 3.

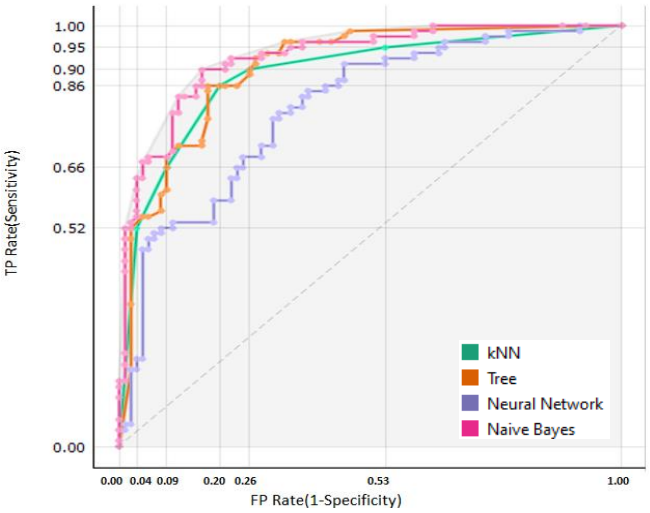


Fig. 3. AUC-ROC Curve.

The figure shows the ROC curves of the tested algorithms and plots the curve for each algorithm, which compares the classification models tested during the study. The curve shows a false positive rate on the x-axis (1-specificity; the probability that the target = 1 when the true value = 0) versus a true positive rate on the y-axis (sensitivity; the probability that the target = 1 when the true value = 1). The figure shows that the closer the model curve approaches the left boundary and then the upper bound of the ROC area, the higher the accuracy of the classifier/model. Due to the costs of false positives and negatives, the figure can determine the optimal classifier and the threshold for the Naïve Bayes classifier as shown in the figure and, therefore, the highest threshold. Area Under the Curve (AUC), It is clear from the figure that the AUC of the Naïve Bayes (NB) ROC curve is higher than the other classifiers of the KNN, NN, and DT ROC curve. Therefore, we can say that Naïve Bayes did a better job categorizing the positive category in the data set.

F. Discussion and Summary

This research has presented a proposed model to detect fake news on the Twitter platform using machine learning algorithms. The results obtained by applying AI algorithms to our selected features to detect fake news on the Twitter platform show the following results:

- The best Accuracy (CA) achieved by Decision Tree (DT) with 89.5% with these parameters: [minimum number of instances in leaves (10), the smallest subset (5), maximal tree depth (30), the majority reaches (95%)].
- K-Nearest Neighbor KNN achieved the best Precision PR with 90.2% with these parameters: [Number of Neighbors (5), Metric (Euclidean), Weight (Uniform)].
- The best Recall RC achieved by K-Nearest Neighbor KNN and Naïve Bayes NB with 87.837% with these parameters:[Number of Neighbors (5), Metric (Euclidean), Weight (Uniform)] for KNN.
- The best F1-Measure achieved by K-Nearest Neighbor KNN with 89% with these parameters: [Number of Neighbors (5), Metric (Euclidean), Weight (Uniform)].
- K-Nearest Neighbor KNN achieved the best Mathieu's Correlation Coefficient MCC with 0.8008.
- We found that when (listed count and followers count) increase, the value of the Target Class is “1” and if the count of (listed count and followers count) is less than 2000, the result of the target Class is "0", with considering the error rate. Table IX shows all results achieved in the Evaluation/confusion matrix and MCC Results.

TABLE IX. EVALUATION RESULTS

	TP	FP	FN	TN	CA	PR	RC	F1	MCC
KNN	65	7	9	81	0.89	0.90	0.88	0.89	0.80
DT	63	7	11	81	0.89	0.9,	0.85	0.88	0.78
NB	65	11	9	77	0.86	0.85	0.88	0.87	0.75
ANN	62	18	12	70	0.82	0.78	0.84	0.81	0.63

The results in Table IX show the following:

KNN: The proposed model was able to find out the following:

- Predict 65 truthful tweets and 81 false tweets.
- Failed to predict seven truthful tweets and nine false tweets.
- Predicting the correct news with a precision of 0.90.
- Predicting the incorrect one (Recall) of 0.87 MCC was 0.8, and this value near ($^{+}$1) means perfect accuracy.
- The KNN Evaluation result was the best.

DT: The proposed model was able to find out the following:

- Predict 63 truthful tweets and 81 false tweets.
- Failed to predict seven truthful tweets and 11 false tweets.
- Predicting the correct news with a precision of 0.90.
- Predicting the incorrect one (Recall) of 0.85
- MCC was 0.7, and this value near (+1) means good accuracy.

NB: The proposed model was able to find out the following:

- Predict 65 truthful tweets and 70 false tweets.
- Failed to predict 11 truthful tweets and nine false tweets.
- Predicting the correct news with a precision of 0.85.
- Predicting the incorrect one (Recall) of 0.87.

- MCC was 0.7, and this value near (+1) means good accuracy.

ANN: The proposed model was able to find out the following:

- Predict 62 truthful tweets and 70 false tweets.
- Failed to predict 18 truthful tweets and 12 false tweets.
- Predicting the correct news with a precision of 0.77.
- Predicting the incorrect one (Recall) of 0.83.
- MCC was 0.6, and this value was greater than 0.5 and less than +1; this means the ANN achieved the worst accuracy in the proposed model.

In comparison with other approaches presented by other researchers, this approach presents the following:

- Using machine learning algorithms and choosing different attributes (author listed count, author follower count, num of retweets, num of likes). At the same time, other approaches, such as Y. Madani, M. Erritali, and B. Bouikhalene, concluded that the results obtained through machine learning algorithms and Twitter attributes (status count, friends count, follower count) are the same, which we used in our proposed model.
- Using different classifiers like NLP, SVM, RF, and LR, they concluded that the results of machine learning algorithms are better than those obtained from deep learning algorithms.
- The proposed model accuracy achieved was 89.9%, while Y. Madani, M. Erritali, and B. Bouikhalene's present frameworks achieved an accuracy of 79%.
- Y. Madani, M. Erritali, and B. Bouikhalene used Sheryl Mathias and Namrata Jagadeesh's dataset. Still, the researcher collected the dataset and labeled it as the proposed model needs, and the stages were unique, and hard to pick the right tweets.

Finally, the researcher concluded that DT is the best classifier to enhance the detection of fake news on Twitter. The best attributes to enhance the accuracy were author listed count, author follower count, number of tweets, and number of likes.

V. Conclusion and Future Work

This paper presented an improved fake news detection by exploring the methods, techniques, tools, and algorithms used previously on previous proposed models and systems. The authors discussed the classification of fake news related to Covid-19 using machine learning algorithms (ML). The authors concentrated on news on Twitter by enhancing the process of detecting fake news using Machine learning algorithms such as decision tree (DT), Naïve Bayes (NB), K-Nearest Neighbors (KNN) and Artificial Neural Network (ANN) classifiers that can handle and distinguish the real and fake news about Covid-19. Also, the authors proposed a model to detect fake news on a Twitter platform using MLA and meta-date (attributes for a Twitter account). The performance and evaluation of the KNN classifier are the best because the F1 scale recorded by KNN is the highest, the MCC was 0.80%, and the best accuracy for the DT classifier was 0.895%. Moreover, the authors collected correct and fake tweets and corresponding metadata to create a dataset that will be publicly available to other researchers in the same field. In this study, the authors highlighted fake news by applying the proposed model, training, and testing using machine learning algorithms; the authors found that the (Num-Likes, Num-Retweet, Author Followers Count, Author Listed Count) have influenced the results of accuracy and improvement of the effectiveness of the results. Finally, the authors designed an effective and accurate model to detect fake Covid-19 news on Twitter using an MLA using some important features (i.e., Author Followers Count, Author Listed Count). After completing the current research and considering the above results, this research suggests working on text mining to work deeply on the content of the tweet (text) by using, for example, Natural Language Processing and Deep Learning Algorithms, which might be worthy of increasing enhancement of evaluation results. Also, using different languages of Tweets like Arabic and French. Greek. etc., to achieve the best result of evaluation results. On the other hand, the researcher suggests using the largest dataset to enhance the results and applying the proposed model to other platforms and other algorithms techniques.

References

[1] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A Hybrid Deep Model for Fake News Detection," Proc. 2017 ACM Conf. Inf. Knowl. Manag., 2017, DOI: 10.1145/3132847.

[2] K. Shu, D. Mahudeswaran, and H. Liu, "FakeNewsTracker: a tool for fake news collection, detection, and visualization," Comput. Math. Organ. Theory, vol. 25, no. 1, pp. 60–71, Mar. 2019, DOI: 10.1007/S10588-018-09280-3/TABLES/3.

[3] Q. A. Al-Haija, E. Saleh and M. Alnabhan, "Detecting Port Scan Attacks Using Logistic Regression," 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT), 2021, pp. 1-5, doi: 10.1109/ISAECT53699.2021.9668562.

[4] Y. Madani, M. Erritali, and B. Bouikhalene, "Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets," Results Phys., vol. 25, p. 104266, Jun. 2021, DOI: 10.1016/J.RINP.2021.104266.

[5] D. S. Abdelminaam, F. H. Ismail, M. Taha, A. Taha, E. H. Houssein, and A. Nabil, "CoAID-DEEP: An Optimized Intelligent Framework for Automated Detecting COVID-19 Misleading Information on Twitter," IEEE Access, vol. 9, pp. 27840–27867, 2021, DOI: 10.1109/ACCESS.2021.3058066.

[6] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti, and M. Akbar, "Fake news detection using deep learning models: A novel approach," Trans. Emerg. Telecommun. Technol., vol. 31, no. 2, p. e3767, Feb. 2020, DOI: 10.1002/ETT.3767.

[7] G. Kesarkar, S. Babar, P. Aurade, and S. Jaswal, "Hoax News Detection in Twitter," Int. J. Recent Adv. Multidiscip. Top., vol. 2, no. 10, pp. 59–62, Oct. 2021, Accessed: Feb. 02, 2022. [Online]. Available: https://www.journals.resaim.com/ijramt/article/view/1412.

[8] A. R. Mahlous and A. Al-Laith, "Fake News Detection in Arabic Tweets during the COVID-19 Pandemic," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 6, pp. 778–788, 2021, DOI: 10.14569/IJACSA.2021.0120691.

[9] M. S. Al-Rakhami and A. M. Al-Amri, "Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter," IEEE Access, vol. 8, pp. 155961–155970, 2020, DOI: 10.1109/ACCESS.2020.3019600.

[10] L. Alsudias and P. Rayson, "COVID-19 and Arabic Twitter: How can Arab World Governments and Public Health Organizations Learn from Social Media?" Jun. 15, 2020. Accessed: Feb. 02, 2022. [Online]. Available: https://en.wikipedia.org/wiki/Modern.

[11] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," Phys. A Stat. Mech. its Appl., vol. 540, p. 123174, Feb. 2020, DOI: 10.1016/J.PHYSA.2019.123174.

[12] H. Reddy, N. Raj, M. Gala, and A. Basava, "Text-mining-based Fake News Detection Using Ensemble Methods," Int. J. Autom. Comput. 2020 172, vol. 17, no. 2, pp. 210–221, Feb. 2020, DOI: 10.1007/S11633-019-1216-5.

[13] A. M. Sari, N. F. Ariyani, and A. S. Ahmadiyah, "Evaluating the Preliminary Models to Identify Fake News on COVID-19 Tweets," Proc. 2021 13th Int. Conf. Inf. Commun. Technol. Syst. ICTS 2021, pp. 336–341, 2021, DOI: 10.1109/ICTS52701.2021.9607996.

[14] S. Sharma, M. Saraswat, and A. K. Dubey, "Fake news detection using Deep Learning," Commun. Comput. Inf. Sci., vol. 1459 CCIS, pp. 249–259, Sep. 2019, doi: 10.1007/978-3-030-91305-2_19.

[15] P. Patwa et al., "Fighting an Infodemic: COVID-19 Fake News Dataset," Commun. Comput. Inf. Sci., vol. 1402 CCIS, pp. 21–29, Feb. 2021, doi: 10.1007/978-3-030-73696-5_3.

[16] S. Almatarneh, P. Gamallo, B. ALshargabi, Y. Al-Khassawneh, and R. Alzubi, "Comparing Traditional Machine Learning Methods for COVID-19 Fake News," pp. 1–4, 2022, DOI: 10.1109/acit53391.2021.9677453.

[17] A. Yahya, A. Amer, and T. Siddiqui, "Detection of Covid-19 Fake News text data using Random Forest and Decision tree Classifiers," Int. J. Comput. Sci. Inf. Secure., vol. 18, no. 12, 2020, [Online]. Available: https://doi.org/10.5281/zenodo.4427204.

[18] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on TF-IDF and support vector machine for news classification," Proc. 2nd IEEE Int. Conf. Eng. Technol. ICETECH 2016, pp. 112–116, Sep. 2016, doi: 10.1109/ICETECH.2016.7569223.

[19] T. Felber, "Constraint 2021: Machine Learning Models for COVID-19 Fake News Detection Shared Task," Jan. 2021, Accessed: Jan. 31, 2022. [Online]. Available: https://arxiv.org/abs/2101.03717v2.

[20] S. Gilda, "Evaluating machine learning algorithms for fake news detection," IEEE Student Conf. Res. Dev. Inspiring Technol. Humanity. SCOReD 2017 - Proc., vol. 2018-Janua, pp. 110–115, Feb. 2018, DOI: 10.1109/SCORED.2017.8305411.

[21] H. Ahmed, I. Traore, and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10618 LNCS, pp. 127–138, Oct. 2017, DOI: 10.1007/978-3-319-69155-8_9.

[22] M. K. Elhadad, K. F. Li, and F. Gebali, "COVID-19-FAKES: A Twitter (Arabic/English) Dataset for Detecting Misleading Information on COVID-19," Adv. Intell. Syst. Comput., vol. 1263 AISC, pp. 256–268, Aug. 2020, DOI: 10.1007/978-3-030-57796-4_25.

[23] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification," Inf. 2020, Vol. 11, Page 314, vol. 11, no. 6, p. 314, Jun. 2020, DOI: 10.3390/INFO11060314.

[24] O. D. Apuke and B. Omar, "Fake news and COVID-19: modeling the predictors of fake news sharing among social media users," Telemat. Informatics, vol. 56, p. 101475, Jan. 2021, DOI: 10.1016/J.TELE.2020.101475.

[25] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks," Apr. 2020, Accessed: Feb. 02, 2022. [Online]. Available: https://arxiv.org/abs/2004.05861v4.

[26] A. Koirala, "COVID-19 Fake News Classification with Deep Learning," prePrint, no. October, pp. 0–6, 2020.

[27] Albulayhi, K.; Abu Al-Haija, Q.; Alsuhibany, S.A.; Jillepalli, A.A.; Ashrafuzzaman, M.; Sheldon, F.T. IoT Intrusion Detection Using Machine Learning with a Novel High Performing Feature Selection Method. Appl. Sci. 2022, 12, 5015. https://doi.org/10.3390/app12105015.

[28] J. Miao and L. Niu, "A Survey on Feature Selection," Procedia Comput. Sci., vol. 91, pp. 919–926, Jan. 2016, DOI: 10.1016/J.PROCS.2016.07.111.

[29] N. Hoque, H. A. Ahmed, D. K. Bhattacharyya, and J. K. Kalita, "A Fuzzy Mutual Information-based Feature Selection Method for Classification," Fuzzy Inf. Eng., vol. 8, no. 3, pp. 355–384, 2016, DOI: 10.1016/j.fiae.2016.09.004.

[30] Abu Al - Haija Q, Al Badawi A, Bojja GR. Boost - Defence for resilient IoT networks: A head - to - toe approach. Expert Systems. 2022 Jan 3:e12934.

[31] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proceedings, Tweet. Int. Conf. Mach. Learn., vol. 2, pp. 856–863, 2003.

[32] "scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation." https://scikit-learn.org/stable/ (accessed Feb. 02, 2022).

[33] Al-Haija, Q. Abu, and A. Ishtaiwia. "Machine Learning Based Model to Identify Firewall Decisions to Improve Cyber-Defense." International Journal on Advanced Science, Engineering and Information Technology 11, no. 4 (2021): 1688-1695.

[34] Alnabhan, M., Habboush, A. K., Abu Al-Haija, Q., Mohanty, A. K., Pattnaik, S., & Pattanayak, B. K. (2022). Hyper-Tuned CNN Using EVO Technique for Efficient Biomedical Image Classification. Mobile Information Systems, 2022.

[35] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) are more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," BioData Min., vol. 14, no. 1, pp. 1–22, Feb. 2021, DOI: 10.1186/S13040-021-00244-Z/TABLES/5.

A 4-Layered Plan-Driven Model (4LPdM) to Improve Software Development

Kamal Uddin Sarker[1]
Department of Computer Science
American International University Bangladesh
Dhaka, Bangladesh

Aziz Bin Deraman[2]
Faculty of Ocean Engineering Technology and Informatics
University Malaysia Terengganu
Terengganu, Malaysia

Raza Hasan[3]
Department of Computing and IT
Global College of Engineering and Technology
Muscat, Oman

Ali Abbas[4]
Department of Computing
Middle East College
Muscat, Oman

***Abstract*—Quality is the degree of excellence of a product and one of the most important factors of software projects that mainly defines user satisfaction and success of the project. Software methodologies represent a variety of tasks, processes, and roles to manage time, cost, and quality. The invention, innovation, and diffusion for technological advancement creates challenges of software projects, thus several existing methodologies albeit with limited scope. A software product is highly influenced by the latest technology and distributed project management opportunities. Management issues are introduced for a virtual project management environment when resource persons are in another corner of the world. To resolve the problem, this research presents a new software project management model (4-LPdM) with alternative actions and practices to effectively manage. The model was presented to 20 different organizations and 29 respondents gave feedback who had experience between 1-16 years in multiple sections of software engineering. The model is evaluated based on the factors of advanced PMBOK 4.0 (scope, cost, quality, resource, risk, plan) and two (management, sustainability) additional features according to the demand of experts. This research illustrates statistical analyses to examine the significance of the proposed model besides a comprehensive comparative study of the traditional methodology.**

***Keywords*—Software development methodology; project management; 4-layered plan-driven model; quality factors; sustainability**

I. Introduction

Software engineering is continuously upgrading with its area of application and research with the advancement of technology by accepting new opportunities and overcoming challenges. The capability and features of the software are increasing with efficient data handling and adding smart functionalities. Smart devices, industrial automation, artificial intelligence equipment, and digital business concepts enhance opportunities for software industries and researchers. An industry expert or researcher can work from any corner of the world through a distributed working environment. Software engineering process management methodologies are improving since the commence of the waterfall model (1956) and followed by various normalized water-fall models (1956-1985), spiral (1986), scrum (1990-1995), rational unified process (RUP, 1996-1998), extreme programming (XP, 1999), agile manifesto (2001), lean (2003), 5th value agile manifesto (2008), DevOps (2009), and Kanban (2010) [1]. Besides methodologies, a set of software quality models is working to improve the quality of the software are McCall's Quality Model (1977), Boehm's Quality Model (1978), IEEE Quality Model (IEEE Std. 729-1983), ISO 9126 Quality Model (1991-2011, ISO/IEC 9126) [2-3]. Plenty of standardization organizations are also working to improve the software project management approach by providing guidelines [4]. But till now 83.9% of information technology (IT) projects completely failed (stop without delivery) or partially failed (compromise with quality) according to Standish Group CHAOS report 2018 [5]. Harvard business review reports that one-sixth of projects run over budget by 200% and 5-15% of projects are failed [6]. Moreover, the project management institute noticed 80% were completed on time without significant wastage of money but the quality is very poor [7] and 75% of IT executives found that their projects were doomed usually from the beginning [8].

So, practitioners, researchers, standardization bodies are working to improve the quality of software products and deduct project failure rates. Digital transformation introduces newer challenges in software industries. Rapidly upgrading technology, increasing functionalities, and changing infrastructure make the software project crucial [9]. Artificial intelligence, cross-platform, Internet of Things, blockchain, continuous development and deployment, progressive web applications, and low-code developments are new trends in software industries [10] that demand more specification of software project information. Near future software engineering becomes close to system engineering that demands structured methodology and architecture for reliable system development with variability features [11]. Furthermore, a distributed software project management system introduces new challenges with diversity culture, different time zone, language barriers, lack of collaboration and communication, trust and ownership of intellectual property, unjustified requirement specification, integration hassle, and lack visionary practice

[12]. A methodology accelerates the project execution process, managing resources, enhancing formal practice, improving sustainability, and ensuring projects quality. Nowadays, systems are improving from on-premises to cloud infrastructure, microcontroller based to industry IoT automation with machine learning, deep learning or reinforcement learning and faces big data challenges (volume, variety, veracity, velocity). Project migration inherits a high grade of complexity with the broader challenge of data collection, specification, sharing, transformation, and analysis when a new technology is being adopted. A well-structured methodology keeps maintainability, portability, and scalability scope with standard guidance and documentation practices.

Current issues in software projects and roles of methodology to address the challenges and make a project successful are mentioned in the introduction. The article is followed by a literature review that consists of reasons for software project failure and the importance of upgrading methodologies. Section III proposes a methodology that is illustrated by Fig. 1 and elaborates on its functionalities in the sub sections. The research methodology discusses ways of evaluation of the proposed methodology. The proposed methodology brings quality of process and product in an architecture. It consists of process, task, and people under the quality control framework. Section V discusses detailed outcome of the study and concludes in section VI with limitation, future work and remarks of contribution.

II. Literature Review

Project management approaches are improving by introducing new inventions in process, tools, and management. Software process management methodologies involve the invention to complete projects effectively. Software development endeavors are affiliated with the practice of a creative project management approach. Software projects are regularly related to innovation in management to overcome the challenges of the fast and dynamic changing of technologies and focus on the best services and products. A higher degree of creativity and flexibility are required in practice with the innovative process of methodology. This section consists of short literature on the reasons for project failure, the role and progress of methodology, and the existing gap.

A. Software Project Failure and Methodology

A project executes by a group of team members with distinguished responsibility that might be varied from organization to organization and project to project based on the type and nature of the project, mission and vision of the organization, and business goals. Potential stakeholders work in a team to make a project successful, but a significant number of projects fail to include its' users. A project commences with a determination to complete on time and budget, but it faces difficulties in the execution period with factors related to process, task or person. As a result, it becomes a challenging project and if the project team is unable to overcome issues it will fail. A project has a chance to face challenges by the wrong strategy of a project or organization [13], wrong or unrealistic planning [14], lack of stakeholders' support [15], and weakness of project management professionalism [16]. Discenza and Forman show the importance of adequate technical and non-technical resources, maintaining scarce resources, promoting effective communication, utilization of technical tools, and managing stakeholders' decisions by using operational metrics to make a project successful [17]. Kulish noticed that complexity of the design and code linearly related to the number of errors in a product. Time constraints of the project, human intervention factors, and miscommunication enhance projects' complexity [18]. Reasons for project failure are associated with people, technology, process, company, leadership, and business goals [19]. Uncertainty or risk is one of the most important reasons which makes a project fail [20] and it appears from stakeholders, technology, or nature. A model and methodology can set roles to develop an effective personality of stakeholders that could contribute to reducing the risk of a project [21]. A methodology approaches a systematic workflow and control of the project. It views on justified requirements, helps to estimate logical cost and effective hours, guides to incorporate change management, keeps standard documentation, ensures tracking on functional review, monitors, and controls on the project, allows backtracking if need (few cases), encourages formal communication among stakeholders, and helps to measure the size of the team of potential stakeholders [22].

B. Commonly used Methodologies

Software process management methodologies have distinguished features and each one has special contributions in software engineering. The waterfall model is the first formal and most influential in software engineering [23]. It has sequential logical phases where one phase accepts feedback for the previous stage. The fundamental waterfall model is modified by overlapping functionalities of phases to utilize time and resource effectively [24]. But it is rigid with fixing requirements and confirming documentation at the earlier state of the project; moreover, users can share their suggestions only at the beginning. So, it is not appropriate for the projects where the requirements can change after execution of a project. The incremental approach is applied in software project management to bring more flexibility, where the client gets a solution part by part and the user can give feedback until the end of the project; iterative scope allows to give a partial solution and it should be updated by the several numbers of iterative feedback from users [25]. But iterative and incremental approaches have no standard architecture, so it is difficult to update and maintain the software. Too much user interaction in an iterative approach increases the scope of arguments. For example, the Spiral model is an iterative approach that is appropriate for high-risk and complex projects but difficult to implement time and cost constraint projects [26]. V-model integrates testing in all phases of the model to ensure the quality of the product [27], but not suitable for high risk, complex, object-oriented, and the project with moderate requirements. Parallel processing is initiated by the Rational Unified Processing model (RUP) which is time constraints iterative system, but it only focuses on functional requirements [28]. These are called heavyweight documentation-oriented plan-driven methodologies.

The agile approach brings innovation in software project management that helps to complete a task on time and does not support heavyweight documentation practice [29]. Agile

methodologies are adopting a heterogeneous number of dynamic software projects where an organization's environment changes rapidly [30]. Agile is suitable for an organization that has a high probability to change management policy-procedure, tools and techniques, and working environment [30]. The agile manifesto is the foundation of agile families and scrum is popular in the agile family and it helps to manage complex projects by integrating creativity [31-32]. Extreme Programming (XP) allows customer interaction that operates by short iteration, Cristal methodology tailors' business goals and Agile Software Process (ASP) supports faster development [33]. Kanban method emphasizes business agility and realistic planning to deliver software products just-in-time [34]. Dynamic system development methodology (DSDM) emphasizes quality products in time constraints with limited iteration [35].

Build and fix is a methodology with lack of architecture and formal feedback which is reactive, and problems are fixed only when they occur. Waterfall is a liner approach where each phase is completed before continuing another one and there is lack in formal change management as well as feedback collection before completion of a project. V-shape is like waterfall but more concentrated on verification and validation in each phase and ignored risk analysis. Prototyping consists of three variations named rapid prototyping used for testing, evolutionary prototyping used for incremental improvements of the design, and operational prototyping improves the speed of production. Incremental consists of multiple cycle of development where entire process can restart any time that allows to change requirements and update a system. Spiral introduces risk analysis in iteration where new requirements are funneled and allows testing earlier. Agile is an umbrella of multiple methodologies that focuses on efficient and iterative development in an agile team.

Software methodologies could be divided into two major categories: plan-driven, and agile. Plan-driven are heavily weighted with documentation and rigid with a systematic approach. On the other side, agile methodologies are light weighted and have time constraints. Both approaches have pros and cons, such as plan-driven is process-oriented and it does not support requirement change frequently while agile methodologies face problems in maintenance and upgrading of a product.

Project management activities consist of methods, tools (e.g. Gantt Chart, network diagram, work breakdown structure), software (Microsoft Project), decision-making methods like feasibility study, risk analysis, and communication plan with collaborative tools: video conferencing [36]. The recent study (2019) of Walker and Lloyd noticed that project management work would be positive for non-routine workers by accepting advanced technology in the 2030s because of the border-free distributed working space [37].

C. Current Project Management Issues

Unclear scope, time constraint, requirement changes, poor communications, managerial weakness, lack of formal practices, unrealistic resource allocation and planning, and insufficient testing are the common issues in software project management. But due to the technological advancement new challenges are appearing for AI, IoT, and big data projects. Technological projects need information specification for accuracy, reusability, scalability, and maintainability. Artificial intelligence applications, IoT software, and big data platforms use huge information that need to be specified explicitly by concept, role, and axiom [38-39] and descriptive logical or ontological presentation that improves ambiguity-free information for a shared domain [40]. Sarker et al. proposed a structure to develop and practice own methodology to consider effective internal and external stakeholders' participation for each project based on the user requirements and business goals [41]. Explicit information specification with ontology, descriptive logic, graphical presentation can reduce complexity of a project and improve communication among stakeholders [46]. A monolithically presented methodology with controlled language use to generalize the process of the methodology and improve integrated performance [44]. A methodology should consider sustainability factors into the product and encourages sustainability practices in project implementation [45].

Project management tools support to manage a project virtually and resource person can be distributed to the world that reduce office management cost, access talent from any corner of the world, and increase productivity; but need to overcome the challenges like virtual monitoring, multicultural team, trust on distrusted employee, and virtual communication [50].

4-LPdM proposed a formal approach, well defined framework, focus on information specification and four layered quality assurance that will reduce the issues of AI, big data, IoT, and distributed projects besides resolving regular issues.

III. Proposed Model

4-Layered Plan-driven Model (4-LPdM) (Fig. 1) distributes the tasks into phases that are arranged in a logical order of waterfall architecture, but four transitions are specified called the layers. The first layer consists of requirement analysis and scheduling of the project, the second layer for in-depth design purposes, the third layer consists of coding and testing for both unit and system, while the fourth layer performs formal closing of the project. Furthermore, it shows the importance to specify the stakeholder, task, tools-techniques for each layer that helps to guide the model.

A. Phase-1: Requirement Specification

The first layer (Fig. 1) consists of requirement collection and analysis from the respective sources where the users, experts, manager, and system analyst are the main key persons to accomplish the tasks. A user can share the visible requirements of functions that are required for the system. An expert will justify the requirements based on the demand of the market and competitors' values. A system analyst can support a manager to make decision for requirement fixing, technology selection, and cost benefit analysis. The 4-LPdM shows interest in recognizing, defining, measuring, and analyzing the requirements to sustain, improve, monitor and control. This phase maps business functions to the software process. Moreover, the planning phase has plenty of tasks that guide monitoring and control of the project. It suggests the utilization

of tools, applications, and techniques to monitor, track, and control a project. This phase asks to develop stakeholder management policy, feedback accepting procedure, and communication plan. A good plan should consist of concrete goals, milestones, and tasks that are specified by date and tracking number. It also includes cost, time, and resource allocation for each task. Risk management activities are included to identify and take mitigation plans at the beginning of the project. Moreover, the feasibility study will help to measure the outcome of the project concerning customer/user requirements (functional and non-functional) and business goals of the vendor and client by cost-benefit analysis. The technical, operational, and ethical feasibility study will improve the acceptance of the system. This layer sets actions to fulfill the vision and mission of the project and achieve business goals. The quality team of this layer will review quality factors so that manager could incorporate required quality functions and information specification to lead IoT, big data, AI and distributed projects.

B. Phase-2: System Design

Designers design the interface by incorporating accessibility and usability requirements; data design consists of standard specification, convention and controlled language for the project and management information; efficient databases consists of normalized tables with required integrity and constraints; control language (algorithm, pseudocode, descriptive logic, predicate logic for reducing misunderstanding), diagrams (context diagram, data flow diagram, entity-relationship diagram, sequence diagram, flowchart, etc.), interface, etc. In this phase (Fig. 1), feedback is expected from users, experts, and system analysts to ensure completeness and quality of the system. It keeps scope of the interaction of potential stakeholders. The design phase becomes more flexible than the traditional waterfall model because until the finalization of the design the user can change requirements. Customer interaction and satisfaction are extremely important for approving the design and passing the phase. System analysts can clarify requirements if required to the designers and they will finalize the architecture of the software. This phase concluded with structured documentation of earlier stages that could be shared through a distributed system, controlled by the manager, and flexible for reusability. This phase is more important to ambiguous free information specification for big data, IoT, AI projects and how to ensure effective communication in virtual project management.

C. Phase-3: Development and Deployment

The third layer (Fig. 1) corresponds to the development, testing, and deployment phase which starts with coding. Effectiveness of the development and testing of a project is dependent on the quality of the design and explicit specification of data is mandatory for virtual or distributed project. The approved design of the previous phase is transformed into a programming notation according to a computer language. Programmers can easily convert the controlled language expressions of documentation and design materials. A standard notation can be easily converted to a program and support for test case generation. Unit testing is simultaneously performed by the quality control unit and developers. Experts are suggested to do template-based testing for accuracy and efficiency. A module consists of related units that are integrated and again tested by experts. The system is tested by quality control before handover to the customer, as well as after deployment, a short time of user training and testing is suggested so that users can use and manage the software effectively. This methodology focuses more activities in previous three phases to reduce complexity of fourth phase. Our proposed methodology supports to keep stand documentation that will improve maintainability, scalability, and portability for a project.

D. Phase-4: Formal Closing

If the deployment test is satisfactory, then only it can be released for operation and the project enters in closing state (Fig. 1). A project teaches a lot of lessons to the stakeholders. Especially, mainstream project execution members get good experience related to technology, management, and communication. The closing layer of the project guides to analyze recently passed experiences so that team members can enhance their potential for future projects. In addition, it improves archiving quality to ensure reusability for further requirements; it helps to update project level or organization level policy procedure, and it is a good motivation for team members. 4-LPdM keeps formal documentation for reusability, maintainability, and scalability for an existing project, but it could be reuse for similar types of more future project and improve sustainability practice too. Project assessment and strategic planning will be easier for formal documentation practice.

E. Capability Study of 4-LPdM

The proposed model is developed based on the demand of the near-future software project highly influenced by artificial intelligence, big data processing, embedded system, and significant risk for technological change. These systems require highly specified information by role, axiom, and concept. Moreover, there is a scope to hire (virtual environment) global talents for the project software projects. This section consists of the following capabilities of 4-LPdM.

F. Project Management Capability

The 4-LPdM specifies the smallest unit of tasks for accurately measuring the size, functional dependencies, and complexity of the project. The project manager and his team define the errand and the execution team performs accordingly. The model prescribes to maintain standardization guidelines and values. Concerning requirements gathering, it gives importance to the client, system analyst, and experts from business and market domains. The design is developed based on the analysis of requirements that stop the scope creep possibility. Time, cost, and assets allocation ensure effective utilization at the micro-level. Critical risk mitigation and communication plans improve the awareness and responsiveness of the stakeholders. 4-LPdM blends the system development model and project management to improve the quality of products and processes. It gives the importance of utilizing tools and techniques for effort estimation, realistic plan development, stakeholder management, and formal documentation practice that will ensure an effective management process.

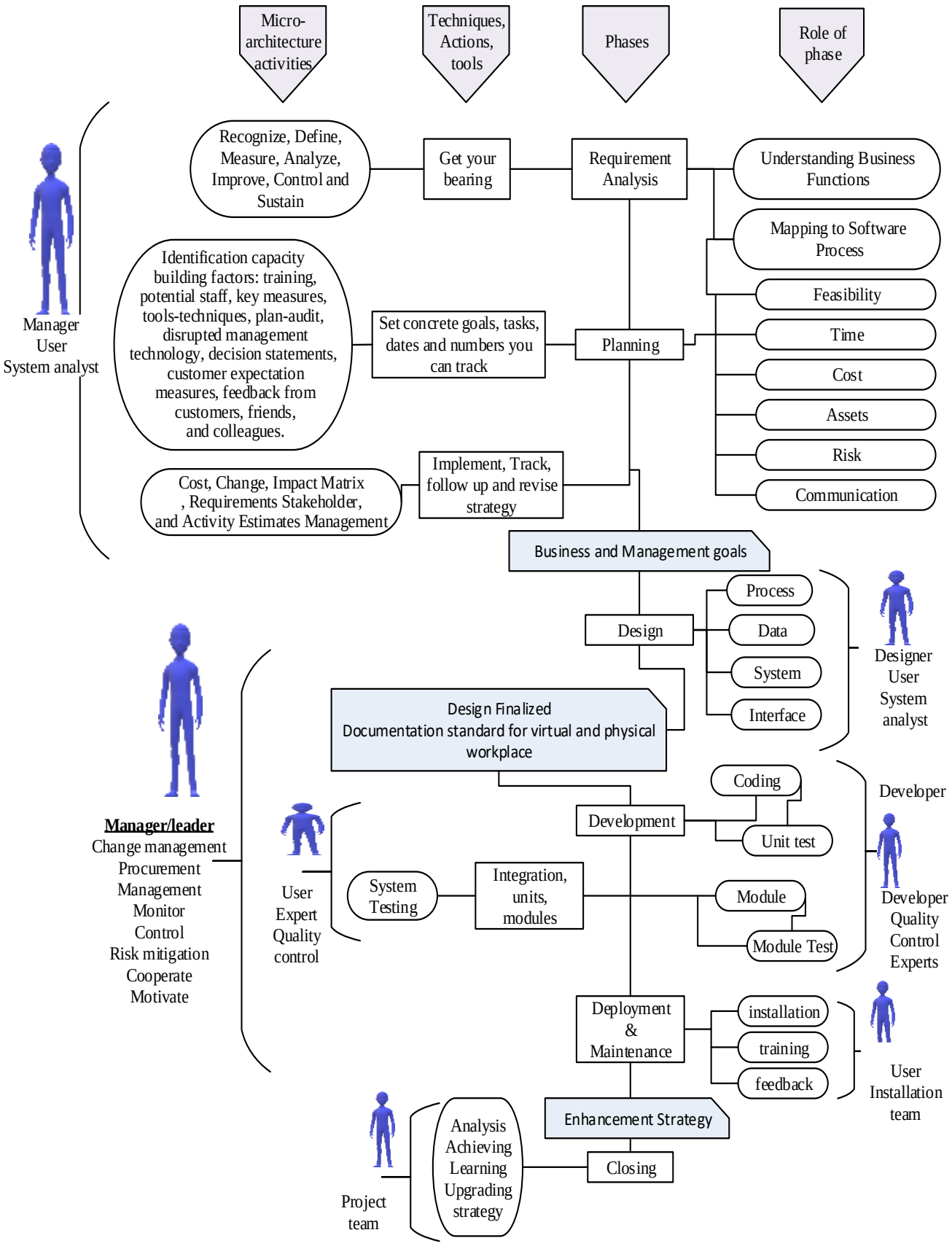


Fig. 1. 4-Layed Plan-Driven Model (4-LPdM).

G. Timeframe for a Phase

4-LPdM specified four major layers instead of phases of traditional methodologies due to incorporate project management capabilities; and each phase is specified with actions, process and respective stakeholders but not fixed the timeframe that depends on the size and complexity of the functionalities and resource availabilities. External stakeholders' requirements and business needs also influence the calculation of time. 4-LPdM suggests that a manager should consider additional time for meetings, communication, and analysis for each level besides resource specification. Moreover, it shows importance to implement project management tools and techniques to calculate, visualize, and manage schedule.

H. 4-LPdM and Quality Control

4-LPdM is highly visible to present the task with the designed required stakeholder, tool, technique, and process that should be maintained by the execution team. The visibility features of the methodology help to determine the realistic time, cost, and resources to improve quality. Customer or user involvement significantly impacts the methodologies, and it specified the purpose of their interaction for a particular task or process. It allows customer interaction more than the heavy-weighted traditional methodologies and reduces over-interaction of light-weighted methodologies, so it ensures reasonable and justified customer interaction that improves customer satisfaction. The model proposed for specialized project management of near-future software development with artificial intelligence, smart infrastructure with IoT, and big data processing for knowledge retrieval. So, it explicitly specifies each task, process, and domain information on standard documentation (e.g., controlled language) to reduce ambiguity which enhances project management capability and quality.

Traditionally time, cost, and scope are considered the most significant influential factors of a software project to maintain the quality; that is updated by the project management body of knowledge (PMBOK 4.0) with six factors: scope, budget, quality, schedule, risk, resources [42] [44]. Mohammed et al [43] developed a six-pointed star model to evaluate the effectiveness of their proposed model with factors time, product (scope), risk, cost, and resource [43] [45]. Customers want to get a product on time within their budget and that should carry all functionalities so quality is described by customers' satisfaction and business goal of the organization. Fig. 2 is the quality model that consists of four factors to ensure the quality of the project based on the traditional model of scope, cost, and time. The project team considers a project as a successful project when the customer is satisfied with functional and non-functional requirements within schedule, budget, and scope. The task of a project is executed by a systematic process that should be well documented according to standardization guidelines. Effective resource allocation reduces extra cost, and an appropriate tool accomplishes a process on time without compromising the quality of the product. Stakeholders are guided by the model to practice formal documentation and responsible resource utilization that improve the efficiency of the workflow. The comprehensive and concrete quality control model is proposed (Fig. 2) to guide the execution of 4-LPdM. 4-LPdM is the main contribution of this research that will guide software practitioner to enhance the organizations' quality in software project management and improve quality of the product. This model inspires to practice guidelines of standardization organizations but not recommend for certification (individual choice). It is also suggested for standard organizations that has capability of utilizing tools and technologies.

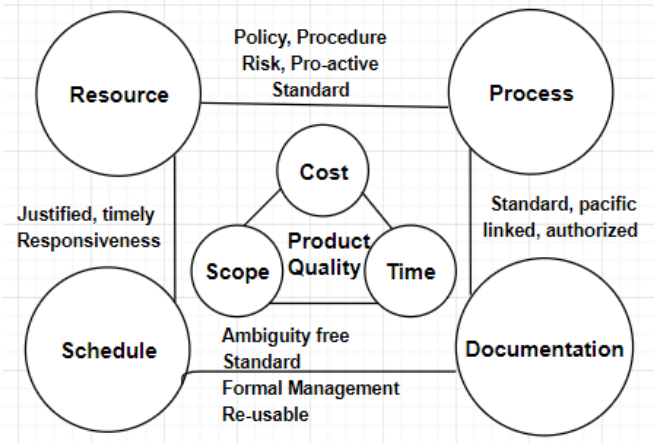


Fig. 2. Quality Control by 4-LPdM.

IV. Research Methods

This section consists of data collection and analysis. Individual expert opinions are collected to determine the validity and efficiency of the proposed model. Furthermore, the proposed model is presented to the experts before data collection. Scope, cost, time, quality, sustainability, risk, resource, and management are considered influential factors according to the suggestion of experts of software developing companies. According to the project management body of knowledge (PMBOK 4.0), the advanced model of triple constraints consists of quality, scope, resource, budget, risk, and schedule known as influential factors of software project management. But management is one of the most vital factors to make a project successful; and sustainability improves quality by reducing wastage of resources: time, cost; and assets. So, management and sustainability are the additional two factors that are considered for the evaluation of this model. Software Requirement Specification (SRS) and milestone of a project are controlled by scope factors; a realistic plan is developed and executed with schedule factors; the budget parameters are justified by the return of the investment; resource factors are used to ensure efficient utilization of assets; sustainability factors improve the quality of process and product, while overall satisfaction is measured by the quality factors. Furthermore, complexity, understandability, and appropriateness are three more criteria that are considered for general reflection. Validity and efficiency checking of the proposed model is the main aim of the identification of the aforementioned factors.

For evaluation, a set of well-known software firms are invited from Bangladesh and abroad who have local and/or international experience. 29 representatives from 20 organizations were accepted to attend the evaluation process who were trained by poster and online presentation. They study

the proposed methodology and try to implement their project (exiting / new project) for a month and evaluate based on the findings. Participants are selected from different levels of experience manager, developer, lead developer, software engineers, and system analysts, and more than 50% have multiple levels of experience including free lunching, individual, and teamwork.

A survey was conducted to collect feedback from experienced people of software firms. 29 respondents give feedback who are from 20 different organizations and 3 of them have freelancing experience. There were two different sections in the questionnaire: i) respondent and his/her organization's information was in section-1 and ii) section-2 carries responses for the proposed model. Table I (a) shows the respondents' experience in software production. Table I (b) describes the mostly practicing methodologies of the organizations. Survey respondents were related to all phases of the software development life cycle. They had different experiences on different types of projects. System analysts, designers, requirement engineers, developers, managers, testers, marketers are the common types of respondents. Respondents' current position in their organization is described in Table I (c). Table I (d) illustrates the respondent's professional experience.

TABLE I. RESPONDENTS' ANALYSIS

(a) Production Classification		
Respondent's Experience	**Frequency**	**Percentage**
Local Production	10	34.48%
International Product	11	37.93%
In house Product	4	13.79%
Local & Global Product	4	13.79%
(b) Practicing Methodology		
Respondent's Experience	**Frequency**	**Percentage**
Waterfall Methodology	3	10.34%
Agile Methodologies	15	51.72%
PRINCE2 Methodology	2	6.90%
Self-developed Methodology	7	24.14%
Other Methodologies	2	6.90%
(c) Respondent Position		
Respondent's Experience	**Frequency**	**Percentage**
Developer	5	17.24%
Lead Developer	9	31.03%
Manager	5	17.24%
System Analyst	4	13.79%
Software Engineer	6	20.69%
(d) Respondents' Experience (Years)		
Respondent's Experience	**Frequency**	**Percentage**
1-4 Years	8	27.59%
5-8 Years	13	44.83%
9-15 Years	6	20.69%
More than 15 Years	2	6.9%

V. EVALUATION

This section illustrates the statistical analyses of the collected numerical responses. These analyses aim to show the influence of each factor and how the management is related to each other. Table II shows the statistical analysis that compares eight factors in the form of Relatively-Importance. This methodology supports resource management (0.8344828) mostly and least interest in cost management (0.7609195). Table II is summarized from the data analysis of Appendix A. There is not much variation among the eight measuring factors (standard deviation). Appendix A describes the summarized result of collected responses that consists of all achieved frequency of 29 participants. Total frequency and computed percentage weight are represented according to the Likert scale. The significance of each factor of the proposed model is reflected in Appendix A, hence it shows that "strongly disagree" and "disagree" are too much less than comparatively "strongly agree" and "agree". Hence, only the "strongly agree", "agree", and "neutral" frequency table is illustrated in Fig. 3.

The average score for all factors is in between 3 and 4 of Likert scale (strongly agree=5, agree=4, neutral=3, disagree=2, strongly disagree=1). Fig. 3 represents responses for each sub-category (for example scope has three subcategories (A, B, C) and satisfactory is comparative more than any other options. The average score of all sub-categories is the final score for each category. 3.8 to 4.2 are the average score of the proposed model that is very near to agree (4) on the Likert scale. For example, the Likert scale value is multiplied with the average value of each factor then again calculate the average for all responses of this domain (scope, plan, etc.). Thus, average score represents the positive feedback in all aspects with minor variation. Standard deviation from 2.4 to 2.8 and according to the empirical role, it shows more than 95% response lies beside the means (i-1 to i+1). Relative importance is calculated (R_i=1/5N_i ($5n_5$+$4n_4$ +$3n_3$+$2n_2$+$1n_1$)) and presented in Table II to show the importance of factors. The values are very close to each other, and the range is 0.7609195 to 0.8344828 indicates correlation.

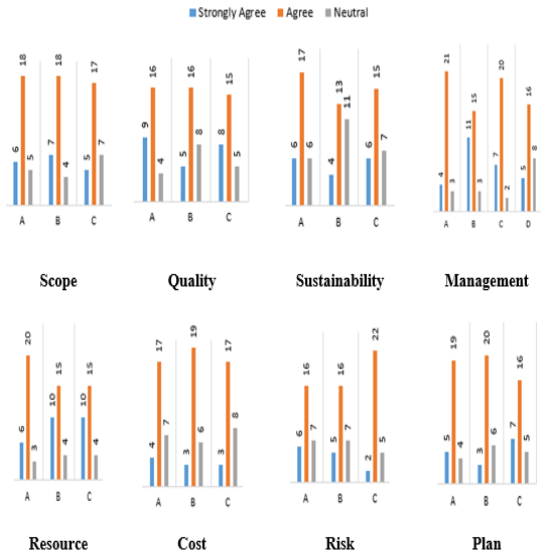


Fig. 3. Frequency Graph for All Measuring Factors.

TABLE II. STATISTICAL ANALYSIS OF THE 8 FACTORS

Factor	Mean	Standard Deviation	Relatively Importance
Scope	4.1944828	2.7620831	0.8045977
Quality	4.0344828	2.4832522	0.8068966
Sustainability	3.862069	2.410228	0.7724138
Management	4.0775862	2.8300027	0.8155172
Resources	4.1724138	2.699799	0.8344828
Cost	3.8045977	2.7849991	0.7609195
Risk	3.908046	2.8326754	0.7816092
Plan	3.954023	2.8490392	0.7908046

TABLE III. OVERALL SATISFACTION

4-LPdM Implementation		4-LPdM Understanding		Appropriate for Projects	
Complex	20.7%	Easy	31%	Medium to large	72.4%
As usual	55.2%	Acceptable	58.6%	Small to medium	13.8%
Simple	24.1%	Difficult	10.4%	All	13.8%

Therefore, the relative importance ranks indicate positively correlated with other factors. Table III shows general aspects of three measures of complexity of the model, understanding the functionality of the model, and when it is suitable. The average percentage (55.2%) of respondents found that it is as usual as others in terms of complexity. 20.7% considered the presentation of the methodology is complex while 24.1% feels it is simple for them. The respondent understands from the presentation and poster and after implementation their project 10.4% feels difficult due to the explicit information specification. 31% feel easy to understand for implementation but more than 58% recommended as acceptable. This methodology is recommended for medium to large projects (72.4% in Table III) because of the extra activities for information specification that will accelerate cost for the small projects.

A. Comprehensive Comparative Study

Waterfall involves users and customers only at the initial stage of the project, so it freezes requirements and documentation at the first phase. It also faces uncertainty problems and measuring the progress of the project is difficult too. The proposed model allows customer interaction and requirement flexibility until the design is finalized as well as risk and quality management mitigate uncertainty. The agile methodologies aim to accomplish a project in a short time that could compromise with quality, and lack of documentation practice makes problems in the re-usability of design and code. Furthermore, new employees struggle in an agile team for technology transfer and highly functional dependency projects. A well-structured and documentation practice of the proposed model demolishes the issues of agile methodologies. Spiral is good for high-risk, complex, and without time constraint projects but time and budget are crucial factors of any project, but the proposed model is suitable for medium to large projects with time constraints. An iterative approach does not fix requirements at the early stage that may cause ambiguous requirement specifications, it allows more customer interaction and informal practice that could make problems to accomplish a project in time and budget. The proposed model allows user interaction only before fixing the design. In addition, its' formal practice of communication and documentation addresses the limitation of the iterative model.

4-LPdM adopts the plan-driven because more information specification is required for the smart information system, artificial intelligence applications, and big data analytics. Descriptive logic is proposed for documentation due to ambiguity reduction and demolish language barriers in distributed project management; moreover, it can easily convert to computer language. Sustainability is focused to ensure the re-usability of documentation and soft resources for a project to another project. The project management approach is integrated with the system development life cycle to improve the management process, risk reduction, and maximize resource utilization. It is validated by common six factors but also considered other special challenges that appear for new technologies like cloud computing, mobile applications, IoT, AI, and big data project. Table IV describes the role of the proposed methodology for IoT, AI, data science, and distributed projects for current days.

TABLE IV. OVERALL SATISFACTION

Project	Special Features	Role of 4-LPdM
IoT	Connectivity, sensing, scale, dynamic, intelligence, energy, safety, integration [47].	IoT devices should be active and act according to the outcome of intelligence applications. Edge, fog, or cloud computing perform sensing data processing where data processing and sharing is an important activity. 4-LPdM specifies data design (Fig. 1) and suggests formal documentation (Fig. 2) to minimize risk.
AI	Complex training algorithms, computation power, data selection and security, integration, and infrastructure [48].	4-LPdM formal analysis, documentation, and review for each layer; minimize complexity and help to select appropriate algorithms; improve data quality and security; support data interrogation; and manage with a suitable infrastructure. It suggests explicit data specification and design that can improve logic, axioms, grammars, and role fixing for AI applications.
Data Science	Weakness in data literacy and culture, measuring of data science project, stakeholders' cooperation, and trust on solution [39][49].	Stakeholders' cooperation is one of the most important factors in data science to achieve business goals, and (Fig. 1) the proposed methodology includes shareholders in the quality frame (Fig. 2). It also allows explicit data specification [46] and communication roles in the planning phase. It integrates capacity building factors, staff training facilities, measuring tools and techniques, customer interactions, and feedback-accepting systems (Planning phase of Fig. 1) to improve trust.
Distributed	Communication, culture, ownership, misunderstanding, knowledge transfer woes, and hassle of integration [12].	Distributed project management system is new and becoming popular to access low wage technical employees from another part of the world or practicing procurement management for a part of the project. 4-LPdM has a plan-driven approach and recommends using control languages to reduce misunderstanding, support knowledge transfer, and improve communication.
Mobile apps	More interactive with usability and accessibility functionalities [52].	Proposed system will support for interactive, usability, and accessibility features specification in system design phases that will be usable for further similar type projects.
Cloud projects	Public and hybrid clouds project faces migration challenges [51].	Cloud infrastructure become popular due to the pay as you model and small companies migrated to cloud platform. If an on-premises project is completed with 4-LPdM it will be easily migrated to public cloud due to the formal documentation and explicitly project data specification.

VI. CONCLUSION AND LIMITATIONS

A quality project management approach is extremely important to accomplish a successful software project within a predefined time and budget. Furthermore, a product's quality depends on the quality of the process, tasks, and stakeholders that is guided by a methodology. The literature review reveals the importance of mitigating the existing limitations and gaps in software methodologies. Moreover, a methodology should be adaptable and predictable with people, tasks, and processes. The proposed methodology is going to fill up the gap and reduce the limitations by introducing a concrete framework of micromanagement architecture, project management approach, and system development phases. It will enhance the managerial capability by formal process and practice with maximizing stakeholders' responsibility. A quality control unit is adjusted with each stage of the project that will ensure quality and minimize risk.

This model is developed and evaluated according to the opinion of experts and the survey result positively indicates the importance of the proposed model. Statistical analyses (means, standard deviation, relative importance) are applied for scoring the result and positive feedback is reflected in all factors. Therefore, it is appropriate for any standard software developing company.

This is a simplified model that separates virtual management functionalities from the traditional approaches. It overlooked the explanation of traditional phases that will enable an adaptation of existing traditional approaches. The standard software firms that have specific business goals can access talent from any corner of the world. It avoids the complexity of a virtual project management. An ad-hoc or special software could be managed by online procurement management, but it is not logical for a standard organization to host everything in outsourcing without proper utilization of organizational resources. Management software is much more complex with additional functionalities, integration opportunities, monitoring, and control strategies. E.g., this model only performs unit testing in the virtual environment and a user can attend testing in distance mode. While the integrated system is considered into the organization (physical mode). In the future, this work will be extended for a fully virtual mode project management approach.

Limitations: 4-LPdM is a common methodology that is proposed for any software project but considered to resolve latest issues (Section II) too. So, it is evaluated by different types of practitioners (Table I) and based on the six common criteria of general perspective (Table II) to show the overall acceptance for any software project. Number of participants and duration of practice could be increased for further study. It could be extended to the specific software project of AI, IoT, big data and evaluated by the respective experts. Moreover, distributed, or virtual project management approach can implement and evaluate too.

ACKNOWLEDGMENT

The proposed models are evaluated by a group of experts during June-August 2020. We are thankful to the valuable experts who have distinguished experiences in the software industry. Their support is appreciable and makes our research successful. This data is not used for any commercial purposes.

REFERENCES

[1] Henseler, Tomás. (2018). Infographic: Timeline of Software Development Methodologies. Retrieved January 15, 2020, from https://www.hexacta.com/timeline-of-software-development-methodologies/.

[2] P. Nistala, K. V. Nori and R. Reddy, "Software Quality Models: A Systematic Mapping Study," 2019 IEEE/ACM International Conference on Software and System Processes (ICSSP), Montreal, QC, Canada, 2019, pp. 125-134, doi: 10.1109/ICSSP.2019.00025.

[3] AL-Badareen, Anas & Selamat, Mohd & A. Jabar, Marzanah & Din, Jamilah & Turaev, Sherzod. (2011). Software Quality Models: A Comparative Study. Communications in Computer and Information Science. 179. 46-55. 10.1007/978-3-642-22170-5_4.

[4] Standards Organizations. (n.d.). Retrieved March 12, 2020, from https://www.oss.com /resources/standards-organizations.html?gclid=Cj0KCQjwoPL2BRDxA RIsAEMm9y 83GhM3-jQEckajEs1eBDs2aycJqsIt_RcCinfyDbnYypoEwUL7KiQaAuTaEALw _wcB.

[5] The Standish Group report 83.9% of IT projects partially or completely fail. (n.d.). Retrieved April 3, 2020, from https://www.opendoorerp.com/ the-standish-group -report-83-9-of-it -projects- partially-or-completely-fail/.

[6] Bent Flyvbjerg and Alexander Budzier (2011). Why Your IT Project May Be Riskier Than You Think. available: https://hbr.org/2011/09/why-your-it-project-may-be-riskier-than-you-think#comment-section, visited 18th nov 2019.

[7] PMI, 2017. Success Rate Rise, Performing the high cost of low performance. Available:https://www.pmi.org/-/media/pmi/documents/ public/pdf/learning/thought-leadership/pulse/pulse-of-the-profession-2017.pdf visited: 18th Nov 2019.

[8] Geneca, 2017.Why up to 75% of the project will fail? Available:https://www.geneca.com/why-up-to-75-of-software-projects-will-fail/ visited:19 Nov 2019.

[9] Horvath, K. (2018, October 29). How Digital Transformation Impacts Software Development. Retrieved April 3, 2020, from https://content.intland.com/blog/how-digital-transformation-impacts-software-development.

[10] CIKLUM. (2019, September 28). 7 Evolving Trends in Software Development. Retrieved May 7, 2020, from https://www.ciklum.com/blog/7-evolving-trends-in-software-development/.

[11] M. Broy, "Yesterday, Today, and Tomorrow: 50 Years of Software Engineering," in IEEE Software, vol. 35, no. 5, pp. 38-43, September/October 2018, doi: 10.1109/MS.2018.290111138.

[12] ReQtest. (2019, August 6). 7 Project Management Challenges in Distributed Development. Retrieved May 6, 2020, from https://reqtest.com/development/project-management-challenges-distributed-development/.

[13] Cândido, C. J., & Santos, S. P. (2015). Strategy implementation: What is the failure rate?. Journal of Management & Organization, 21(2), 237-262.

[14] Kerzner, H., & Kerzner, H. R. (2017). Project management: a systems approach to planning, scheduling, and controlling. John Wiley & Sons.

[15] Davis, K. (2014). Different stakeholder groups and their perceptions of project success. International journal of project management, 32(2), 189-201.

[16] Kerzner, H., & Kerzner, H. R. (2017). Project management: a systems approach to planning, scheduling, and controlling. John Wiley & Sons.

[17] Discenza, R. & Forman, J. B. (2007). Seven causes of project failure: how to recognize them and how to initiate project recovery. Paper presented at PMI® Global Congress 2007—North America, Atlanta, GA. Newtown Square, PA: Project Management Institute.

[18] Kulish, A. (2019, November 20). Why Quality Software Is Impossible Without Proper Root Cause Analysis (RCA). Retrieved April 20, 2020, from https://www.infopulse.com/blog/why-quality-software-is-impossible-without-proper-root-cause-analysis-rca/.

[19] A Mandal and S. C. Pal. Identifying the Reasons for Software Project Failure and Some of their Proposed Remedial through BRIDGE Process Models.International Journal of Computer Sciences and Engineering. Vol.-3(1), PP(118-126) Feb 2015, E-ISSN: 2347-2693.

[20] Elzamly, A., Hussin, B., & Salleh, N. (2015). Methodologies and Techniques in Software Risk Management Approach for Mitigating Software Failure Risks: A Review. Asian Journal of Mathematics and Computer Research, 2(4), 184-198. Retrieved from http://www.ikprress.org/index.php/AJOMCOR/article/view/63.

[21] A. R. Gilal, J. Jaafar, S. Basri, M. Omar and A. Abro, "Impact of software team composition methodology on the personality preferences of Malaysian students," 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2016, pp. 454-458, doi: 10.1109/ICCOINS.2016.7783258.

[22] Geamba, Cristina & Jianu, Iulia & Jianu, Ionel & Gavrila, Alexandru. (2011). Influence Factors for the Choice of a Software Development Methodology. Journal of Accounting and Management Information Systems. 10. 479-494.

[23] B.Blanchard and W.Fabrycky, Systems Engineering and Analysis, 4th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2011.

[24] Adenowo, Adetokunbo A. A. and Basirat A. Adenowo. “Software Engineering Methodologies: A Review of the Waterfall Model and Object-Oriented Approach.” (2013).

[25] C. Larman and V. R. Basili, ‘‘Iterative and incremental developments. A brief history,’’ Computer, vol. 36, no. 6, pp. 47–56, 2003, doi: 10.1109/MC.2003.1204375.

[26] Dmitry Gurendo. "Software Development Life Cycle (SDLC): Spiral Model". XB Software. October, 2015. Retrieved on 20th November 2020 from https://xbsoftware.com/blog/software-development-life-cycle-spiral-model/.

[27] S. Mathur and S. Malik, ‘‘Advancements in the V-model,’’ Int. J. Comput. Appl., vol. 1, no. 12, pg. 29-34, 2010, doi: 10.5120/266-425.

[28] P. Kruchten, The Rational Unified Process—An Introduction. Reading, MA, USA: Addison-Wesley, 2000.

[29] Birkinshaw, J. (2018). What To expect from agile. MIT Sloan Management Review, 39–42. https://doi.org/10.1142/S201000781100019X.

[30] Baham, C., Hirschheim, R., Calderon, A. A., & Kisekka, V. (2017). An agile methodology for the disaster recovery of Information Systems under catastrophic scenarios. Journal of Management Information Systems, 34(3), 633–663. https://doi.org/10.1080/07421222.2017.1372996.

[31] Varajão, J. E. (2018). A new process for success management. Journal of Modern Project Management, 92–99. https://doi.org/10.1177/004051755602600608.

[32] Stoica, M., Ghilic-Micu, B., Mircea, M., & Uscatu, C. (2016). Analyzing agile development – from waterfall style to Scrumban. Informatica Economică, 20(4), 5–15.

[33] Rajagopalan, S., & Mathew, S. K. (2016). Choice of agile methodologies in software development : A vendor perspective. Journal of International Technology and Information Management, 25(1).

[34] Anderson, D.J. 2010. Kanban: Successful Evolutionary Change for Your Technology Business, Blue Hole Press, Ed.1.

[35] Fahad, Muhammad & Qadri, Salman & Ullah, Dr. Saleem & Husnain, Mujtaba & Qaiser, Rizwan & Qureshi, Shehzad & Ahmed, Waqas & Syed, Shah & Muhammad, Syed. (2017). A Comparative Analysis of DXPRUM and DSDM. International Journal of Computer Science and Network Security, VOL.17 No.5, May 2017.

[36] Jugdev, K., Perkins, D., Fortune, J., White, D. and Walker, D. (2013), "An exploratory study of project success with tools, software and methods", International Journal of Managing Projects in Business, Vol. 6 No. 3, pp. 534-551. https://doi.org/10.1108/IJMPB-08-2012-0051.

[37] Walker, D. and Lloyd-Walker, B. (2019), "The future of the management of projects in the 2030s", International Journal of Managing Projects in Business, Vol. 12 No. 2, pp. 242-266. https://doi.org/10.1108/IJMPB-02-2018-0034.

[38] K. U. Sarker, A. Bin Deraman, R. Hasan, S. Mahmood, A. Abbas and M. Sohail, "Kids’ Smart Campus Ontology to Retrieve Interest," 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), Muscat, Oman, 2019, pp. 1-4. doi: 10.1109/ICBDSC.2019.8645585 available: https://ieeexplore.ieee.org/document/8645585.

[39] Sarker, Kamal Uddin; Deraman, Aziz Bin; Hasan, Raza; Abbas, Ali. Ontological Practice for Big Data Management. International Journal of Computing and Digital Systems. Volume-8, issue-3. Pp:265-272

May2019. DOI: 10.12785/ijcds/080306. Available: https://journal.uob.edu.bh/handle/123456789/3485.

[40] K. U. Sarker, A. B. Deraman and R. Hasan, "Descriptive Logic for Software Engineering Ontology: Aspect Software Quality Control," 2018 4th International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2018, pp.1-5. doi: 10.1109/ICCOINS.2018.8510585 available: https://ieeexplore.ieee.org/document/8510585.

[41] Kamal Uddin Sarker, Dr. Aziz Deraman, Raza Hasan, Abdul Hadi Bhatti, "Software Development Life Cycle Developing Framework" 5th International Conference on Communication and Computer Engineering (ICOCOE 2018) Melaka, Malaysia. https://maltesas.my/msys/explore/pubs/1562833286/1562833286_Article_1564562131.pdf.

[42] K. Schwaber, Agile Project Management With Scrum. Redmond, WA, USA: Microsoft Press, 2004.

[43] MUHAMMAD AZEEM AKBAR, JUN SANG, ARIF ALI KHAN3, FAZAL-E-AMIN4, NASRULLAH, MUHAMMAD SHAFIQ, SHAHID HUSSAIN, HAIBO HU, MANZOOR ELAHI, AND HONG XIANG, "Improving the Quality of Software Development Process by Introducing a New Methodology—AZ-Model". IEEE Access. VOLUME 6, pp.4811-4823,2018, DOI: 10.1109/ACCESS.2017.2787981.

[44] Kamal Uddin Sarker, Aziz Deraman, Raza Hasan, Ali Abbas, Babar Shah, Abrar Ullah. Monolithic Ontological Methodology (MOM): An Effective Software Project Management Approach" Journal of Engineering Research. Vol 10, No 2A, 2022.

[45] Kamal Uddin Sarker, Aziz Bin Deraman, Raza Hasan and Ali Abbas, "SQ-Framework for Improving Sustainability and Quality into Software Product and Process" International Journal of Advanced Computer Science and Applications(IJACSA), issue 9, volume 11, page 69-78. ISSN : 2156-5570 (Online) ISSN : 2158-107X (Print) 2020. http://dx.doi.org/10.14569/ IJACSA.2020.0110909.

[46] Sarker KU, Deraman A, Hasan R, Abbas A. Explicit specification framework to manage software project effectively. Indian Journal of Science and Technology. volume 13, issue 36, page 3785-3800, year 2020. https://doi.org/10.17485/IJST/v13i36.1244 https://indjst.org/articles/explicit-specification-framework-to-manage-software-project-effectively.

[47] Priya Pedamkar. "IoT Features". EDUCBA. 2022. Retrieved on 14th May 2022 from https://www.educba.com/iot-features/.

[48] 10xDS Team. "5 Common Challenges in Implementing Artificial Intelligence (AI)". 10Xds. March 2021. Retrieved on 14th May 2022 from https://10xds.com/blog/challenges-implementing-artificial-intelligence/.

[49] Gramener Inc. "Challenges in Data Science Projects and How to Tackle Them". September 2020. Retrieved on 15th May 2022 from https://blog.gramener.com/data-science-project-challenges/.

[50] Megan Keup. "Virtual Project Management: Benefits, Challenges & Tools". April 2020. Retrieved on 2nd June 2022 from https://www.projectmanager.com/blog/virtual-project-management.

[51] Rushi Patel. 10 Biggest Cloud Computing Challenges in 2022 for IT Service Providers. April 2022. Retrieved on 6th June 2022 from https://www.mindinventory.com/blog/cloud-computing-challenges/.

[52] Radosław Szeja. 14 Biggest Challenges in Mobile App Development in 2022. Netguru. January, 2022 .Retreived on 1st June 2022 from https://www.netguru.com/blog/mobile-app-challenges.

APPENDIX A

Practitioners' Data Analysis

It consists of the analysis of responses of all individual question. There are six major criteria validated by practitioners and each of them divided into 3 to 4 sub criteria to get more accurate reflection. Each question is having criteria for the proposed methodology. For example, the score of resource is the average score of A. how much supportive for hardware/software resource utilization, B. how much suggestive for proper utilization of human resource, and C. how much guided for resource sharing and responsibly handling? There are 29 participants and 6, 20, and 3 are the responses for strongly agree, agree, and neutral; and their percentage are respectively 20.69, 68.97, and 10.34.

TABLE V. OVERALL SATISFACTION

Factor	Questions	Strongly Agree (Response and %)	Agree (Response and %)	Neutral (Response and %)	Disagree (Response and %)	Strongly Disagree (Response and %)	Total and %
Scope	A. The proposed framework will guide to clarify the scope of the project	6	18	5	0	0	29
		20.69	62.07	17.24	0	0	100
	B. It will help to monitor the scope of the project	7	18	4	0	0	29
		24.14	62.07	13.79	0	0	100
	C. It will help to meet the scope	5	17	7	0	0	29
		25	85	35	0	0	100
Quality	A. The proposed framework will help to quality product	9	16	4	0	0	29
		31.03	55.17	13.79	0	0	100
	B. It will improve client satisfaction	5	16	8	0	0	29
		17.24	55.17	27.59	0	0	100
	C. It will suggest a quality working environment	8	15	5	1	0	29
		27.59	51.72	17.24	3.448	0	100
Sustainability	A. The proposed framework will improve economic sustainability	6	17	6	0	0	29
		20.69	58.62	20.69	0	0	100
	B. It will improve social sustainability	4	13	11	1	0	29
		13.79	44.83	37.93	3.448	0	100
	C. It will improve environment sustainability	6	15	7	1	0	29
		20.69	51.72	24.14	3.448	0	100
Management	A. The proposed framework will improve formal management	4	21	3	1	0	29
		13.79	72.41	10.34	3.448	0	100

	B. It will improve professionalism	11	15	3	0	0	29
		37.93	51.72	10.34	0	0	100
	C. It will help to monitor and control	7	20	2	0	0	29
		24.14	68.97	6.897	0	0	100
	D. It will help to distribute project management	5	16	8	0	0	29
		17.24	55.17	27.59	0	0	100
Response	A. The proposed framework will enhance material resource utilization	6	20	3	0	0	29
		20.69	68.97	10.34	0	0	100
	B. It will enhance human resource management	10	15	4	0	0	29
		34.48	51.72	13.79	0	0	100
	C. It will improve cooperation and responsiveness	10	15	4	0	0	29
		34.48	51.72	13.79	0	0	100
Cost	A. The proposed framework will guide to identify cost factors	4	17	7	1	0	29
		13.79	58.62	24.14	3.448	0	100
	B. It will guide to cost control	3	19	6	1	0	29
		10.34	65.52	20.69	3.448	0	100
	C. It will help to complete the project on budget	3	17	8	1	0	29
		10.34	58.62	27.59	3.448	0	100
Risk	A. The proposed framework will help to avoid the risk	6	16	7	0	0	29
		20.69	55.17	24.14	0	0	100
	B. It will help to make the risk mitigation plan	5	16	7	1	0	29
		17.24	55.17	24.14	3.448	0	100
	C. It will focus on business objectives to meet the risk	2	22	5	0	0	29
		6.897	75.86	17.24	0	0	100
Plan	A. The proposed framework will guide to develop a realistic plan	5	19	4	1	0	29
		17.24	65.52	13.79	3.448	0	100
	B. It will help to execute according to plan	3	20	6	0	0	29
		10.34	68.97	20.69	0	0	100
	C. It will help to accomplish on time	7	16	5	1	0	29
		24.14	55.17	17.24	3.448	0	100

A New Model to Detect COVID-19 Coughing and Breathing Sound Symptoms Classification from CQT and Mel Spectrogram Image Representation using Deep Learning

Mohammed Aly[1]*
Department of Artificial Intelligence
Faculty of Computers and Artificial Intelligence,
Egyptian Russian University, Badr City, 11829, Egypt

Nouf Saeed Alotaibi[2]
Department of Computer Science
College of Science, Shaqra University
Shaqra City 11961, Saudi Arabia

***Abstract*—Deep Learning is a relatively new Artificial Intelligence technique that has shown to be extremely effective in a variety of fields. Image categorization and also the identification of artefacts in images are being employed in visual recognition. The goal of this study is to recognize COVID-19 artefacts like cough and also breath noises in signals from real-world situations. The suggested strategy considers two major steps. The first step is a signal-to-image translation that is aided by the Constant-Q Transform (CQT) and a Mel-scale spectrogram method. Next, nine deep transfer models (GoogleNet, ResNet18/34/50/100/101, SqueezeNet, MobileNetv2, and NasNetmobile) are used to extract and also categorise features. The digital audio signal will be represented by the recorded voice. The CQT will transform a time-domain audio input to a frequency-domain signal. To produce a spectrogram, the frequency will really be converted to a log scale as well as the colour dimension will be converted to decibels. To construct a Mel spectrogram, the spectrogram will indeed be translated onto a Mel scale. The dataset contains information from over 1,600 people from all over the world (1185 men as well as 415 women). The suggested DL model takes as input the CQT as well as Mel-scale spectrograms derived from the breathing and coughing tones of patients diagnosed using the coswara-combined dataset. With the better classification performance employing cough sound CQT and a Mel-spectrogram image, the current proposal outperformed the other nine CNN networks. For patients diagnosed, the accuracy, sensitivity, as well as specificity were 98.9%, 97.3%, and 98.1%, respectively. The Resnet18 is the most reliable network for symptomatic patients using cough and breath sounds. When applied to the Coswara dataset, we discovered that the suggested model's accuracy (98.7%) outperforms the state-of-the-art models (85.6%, 72.9%, 87.1%, and 91.4%) according to the SGDM optimizer. Finally, the research is compared to a comparable investigation. The suggested model is more stable and reliable than any present model. Cough and breathing research precision are good enough just to test extrapolation as well as generalization abilities. As a result, sufferers at their headquarters may utilise this novel method as a main screening tool to try and identify COVID-19 by prioritising patients' RT-PCR testing and decreasing the chance of disease transmission.**

***Keywords*—COVID-19; median filter; deep learning; Mel-scale spectrogram; sound classification; constant-Q Transform**

I. Introduction

COVID-19 is an unique SARS disease which first surfaced in 2019 and has since spread over the world, producing a worldwide pandemic [1]. In accordance with the World Health Organization's (WHO) April 2021 report [2] there really are currently over 150 million documented illnesses including over 3 million deaths. Moreover, across over 32.5 million new cases and 500,000 fatalities, the USA has the highest overall number of illnesses as well as deaths.

These large numbers have placed a strain on numerous healthcare systems, particularly given the virus's propensity to cause more genetic variants and spread faster among people.

Recent studies have now employed relatively new artificial intelligence (AI) algorithms to recognise and categorise COVID-19 in CT and X-ray images [3]. Several studies (including CT scans as raw data) used machine and also DL techniques to distinguish among healthy and infected subjects with a discriminating accuracy of much more than 95% [4-9]. The capability of different classifiers including such support vector machines (SVM) and also convolutional neural networks (CNN) to identify COVID-19 in CT images with few which was before stages is the great contribution of these investigations. Additionally, some publications have used DL with supplementary feature fusion approaches as well as entropy-controlled enhancement to detect COVID-19 in CT images [10-13].

In light of the above, this research proposes a thorough deep learning technique for COVID-19 identification using coughing as well as breathing signals (**Fig. 1**). The suggested method might be used as a quick, low-cost, as well as readily distributed COVID-19 pre-screening tool, particularly in locations where the virus has spread rapidly. Despite the fact that the current gold new standard for detecting viral infection, RT-PCR, seems to have a good success rate, it has several drawbacks, such as high costs for equipment and chemical agents, the require for expert medical and nursing staff for tests, breaches of social separation, as well as the long time it needed to achieve outcomes (2-3 days).

*Corresponding Author.

As a result, the construction of a DL model removes the majority of these constraints, resulting in stronger resurrection in the medical and financial domains of many countries.

All of the techniques to sound classification utilise machine learning (ML) and DL. ML classification methods include SVM [14] and decision trees [15], while DL classifiers include CNN models (AlexNet [16], VGGNet [17], GoogleNet [18], ResNet [19]). CNN image classification models are built for speed as well as efficiency. The below are the study's major contributions:

- A novel DL strategy for recognizing COVID-19 from a set of tones.
- The proposed model enhances sound recognition effectiveness by employing a Mel-scale spectrogram and CQT method to convert sound into image.
- Nine DL training models are employed to achieve optimal efficiency.

The present study is structured as follows: Section II of this study examines the existing literature. Section III highlights the major properties of the dataset. Section IV includes a presentation of the proposed COVID-19 cough as well as breath tones model. Section V displays the test findings, whereas Section VI gives the paper's conclusions.

II. Related Work

The rest of the published studies mentioned here used ML and statistical analytics to detect COVID-19 disease. There has been less research that applies CNN and transfer learning on coughing signals datasets to determine the features of normal as well as coronavirus patients. Additional studies on DL with simpler efficiency assessments are thus needed. In this research, a novel model using DL, CQT, and a Mel-scale spectrogram was developed to detect COVID-19. According to the research study here, it is suggested that cough sounds be used to diagnose COVID-19. In fighting the COVID-19 epidemic, the advances are much more efficient and quicker.

This section investigates the much more available research on COVID-19 diagnosis that use coughing signals. As a result, the most recent evaluation of DL for coughing signal scan processing is addressed. This section includes details about the use of ML and also DL in sound detection. The stages of signal categorization can be classified into three phases: pre-processing, extraction, and classification. The core of tone detection study is concentrated on sound generation as well as recognition using classic machine learning approaches [20–22]. This paper concentrated on categorizing and identifying breathing and coughing sounds caused by COVID-19 virus infected individuals. Schuller et al. [23] used CNN to create a deep learning strategy to identify raw breathing as well as coughing in COVID-19 patients. Researchers improved the CNN method, which employs breathing as well as coughing sounds to test whether a person has COVID-19 or is fit. The suggested model is about twice as effective as the typical starting point. The CNN model achieved an overall score of approximately 81%, indicating that a DL model can deliver the best results with the available data.

A CNN model for COVID-19 is shown audio categorization proposed via frequency cepstral coefficients (MFCC) in [24]. The VGG 16 architecture is used in two learning strategies. The provided model achieved an overall of nearly 71 percentage as well as a sensitivity of 81 percentage using a high-quality outcomes method. The authors of Ref. [25] established a methodology for distinguishing COVID-19 and healthy sounds. For training and evaluation they employed 1838 coughs and 3597 other signals divided into 50 groups. In accordance with the study, the DL-based multi-class classifier scored about 92 percent, overall total accuracy. Prior to the COVID-19 pandemic, other research, like [26], to identify cough sound occurrences, a transfer learning technique applied. The NN models are developed in two stages: pre training & fine-tuning, after which the decoded data is collected by a Hidden Markov Model (HMM). In this work, three cough HMMs and one non-cough HMM are included to the proposed model. The experiments were carried out using a dataset generated from twenty two people suffering from various respiratory illnesses. This approach demonstrates that the qualifying deep model can now achieve a 90% precision level. M. aly et al. [27] proposed a classification model to identify COVID-19 in their investigation. The offered dataset contains 1600 wave coughing as well as breathing tones. To convert signal to image, the Mel-scale spectrogram method was utilised. Based on the data, the recommended model's overall accuracy, sensitivity, as well as specificity reached 99.2%, 98.3%, and 97.8%, respectively.

In [28], the author proposed a classification model for pneumonia and asthma. Their approach used MFCC, Shannon entropy, as well as non-Gaussian distributions to quantify signal parameters, and all these attributes have been determined to be the basis for artificial neural network classifiers. The suggested technique has 89 percent sensitivity as well as 100 percent accuracy. The results demonstrate how this technique may be applied to distinguish between pneumonia as well as asthma in public environments. According to [29], the goal of this study is to characterize the unique coughing sounds tones of COVID-19 artefacts in signals from various real-life scenarios. The model provided here tends to take two crucial stages into consideration. Converting the signal to image is the first step, which is improved using the scalogram approach. The second step is feature extraction as well as classification. The dataset utilized comprises 1457 wave coughing tones (755 from COVID-19 as well as 702 from healthy). The machine learning classifier's overall sensitivity and specificity were approximately 94% and 95 %, respectively.

An obvious and common problem with most previous COVID-19 research is that it uses a small dataset. Big data is preferred to little data because the higher the sample size, the more exact your estimations will be. Small data has a few advantages. For example, tiny data makes visualization, examination, as well as knowing what is going on in the data much simpler than enormous data. Furthermore, the innovation of this study is in the development of a DL model based on CQT as well as Mel-scale spectrogram-based breath and cough recordings into this DL model, which performs

much better than conventional respiratory auscultation devices. Where, electronic stethoscopes are preferred because they are more accessible to a larger population. This is essential for obtaining medical data regarding COVID-19 patients in a responsible way, while keeping isolated behavior amongst persons. Moreover, this research examined patients from India, whose COVID-19 has a unique genetic variant likely of eluding the immune system as well as most available immunizations. As a result, it focuses attention on the ability of artificial intelligence algorithms to detect this viral illness in persons with this unique variant, even those who are asymptomatic.

AI design does not need a large amount of memory. This is a good strategy for the future expansion of telehealth and smartphone applications for COVID-19 (or other pandemics) that can offer real-time information and efficient and quick exchanges between patients and healthcare professionals. As a result, as a COVID-19 pre-screening tool, this enables for better and quicker isolation as well as contact tracing than presently existing methodologies.

III. Dataset Characteristics

The dataset for this investigation was obtained from a project aimed at creating an available dataset for pulmonary sounds of normal and unwell patients, which also included participants with COVID-19, according to coswara [30]. Ever since, it has gathered information from over 1,600 people from all around the world (Male: 1185, Female: 415; mostly Indian population). Crowdsourcing was used to gather breathing, coughing, and speech sound using an interactive online app tailored for smartphone devices [31]. All voices were recorded with a smartphone microphone and recorded at a frequency of 48 kHz. All audio samples (in.WAV file) were chosen at random to employ a web interface that enables many writers to review each audio file while also improving labelling performance and accuracy. There are now 120 COVID-19 instances in the database, representing a one-to-ten ratio as compared to normal (control) patients. To produce a balanced dataset, all COVID-19 participants' data was assessed, and the exact number of assessments was assigned randomly from the control participants' data. Furthermore, just two types of breath sounds, shallow and deep, were captured from each patient and used for subsequent research.

The proposed model was developed to classify breathing as well as coughing in order to offer it in a public dataset. This is used by the diagnostic engine. Classifiers for breathing as well as coughing are applied to determine if a sound is connected with COVID-19. We utilized data from the breathing dataset in addition to the COVID-19 and healthy tones dataset to assess the classifier.

IV. Proposed Model

Fig. 1 presents the suggested COVID-19 cough and breath sounds classification model. The architecture design of the proposed Deep Learning cough-breathing classification model is shown in Fig. 2. The presented DL cough classification model needs pre-processing, feature extraction, and classification. The suggested model consists of two major phases. The first phase is feature extraction, which transforms sound to picture using CQT and a Mel-spectrogram, and the second step is feature extraction and classification model. Deep Learning models such as GoogleNet, ResNet18, ResNet34, ResNet50, ResNet100, ResNet101, Mobile-Netv2, NasNetmobile, and SqueezeNet are used in feature extraction and classification. GoogleNet, ResNet, Mobile-Netv, NasNetmobile, and SqueezeNet are the most extensively used Deep Learning transfer learning models. Deep Learning models were employed in the suggested model's learning, validating, and assessing processes for feature extraction and classification.

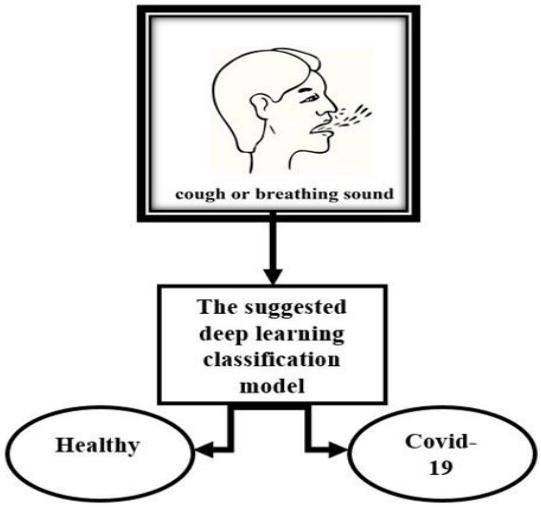


Fig. 1. The Suggested Deep Learning Classification Model.

A. Constant-Q Transform (CQT) and Mel-Scale Spectrogram

Human ears do not register variations across all frequency ranges equally. As frequency increases, it becomes increasingly difficult for individuals to distinguish between separate frequencies. The sound wave will be represented digitally by the voice recording. The CQT transforms a time-domain audio input to a frequency-domain signal [42]. To generate a spectrogram, the frequency will be converted to a log scale, and the amplitude will indeed be converted to decibels (db). The spectrogram will just be translated onto a Mel scale to generate a Mel spectrogram. In order to accurately imitate human ear behaviour using DL models, we employed the Mel scale to quantify frequencies. Each equivalent length among frequencies on the Mel- scale sounds equally distinct to human ears. To transform frequency from Hertz (f) to Mel (m), the Mel-scale utilizes the following equation:

$$m = 2595 \times \log(1 + f/700) \quad (1)$$

A Mel-scale spectrogram is a spectrogram with frequencies estimated in Mel. A Mel-scale spectrogram is a short-time Fourier transform (STFT) value [32]. The CQT and a mel-scale spectrogram are employed in two ways in this work. To decrease noise, the 1-D electrocardiogram (ECG) data will be first standardized. Second, the preprocessed signals are presented to a 2-D mel-scale spectrogram using Continuous Wavelet Transform (CWT). As illustrated in

Fig. 3, the ECG employs CWT to transform the signal from time domain to frequency domain. Convolution using only a median filter utilised to decrease low and high-frequency noise. Small amplitude features of the ECG that are of physiological or clinical importance are generally obscured by noise and interference. Because noise's bandwidth overlaps that of desired signals, basic filtering is insufficient to enhance the signal-to-noise ratio (SNR). The CWT typically uses findings to determine the resemblance of a wave to an evaluation function such as the Fourier transform (FT). The (CWT) is a time-frequency analysis method that differs from the more common (STFT) in that it enables for unlimited high-frequency signal feature localisation in time.

The CWT will accomplish this by using a variable window width proportional to the observer scale—flexibility that will provide for the separation of high-frequency features. The CWT is distinct from the STFT in that it is not limited to using sinusoidal analysing functions. The CWT of function $x(t)$ is measured using equation (2). Where, $\beta(t)$ is father signal, mostly in the time and frequency domains, $\beta(t)$ is a continuous function. (x) is the scale parameter's continuously varying values, and (y) is the position parameter's continuously varying values.

$$CWT(x,y) = (\sqrt{x})^{-1} \int_{-\infty}^{\infty} f(t)\beta(t - y/x)dt \quad (2)$$

The coefficients of CWT coefficients provide a matrix filled with located and scaled wavelets. The father signal's goal is to provide the generation fundamental characteristic of the child signals. Cough tones and breath sounds were separated from the dataset. The COVID-19 and normal groups' symptomatic breathing and coughing sounds were compared over time to see whether there had been any differences (Fig. 3).

B. Deep Learning Models

Many successful pre-train CNNs are capable of passing learning. Furthermore, they require dataset preparation and analysis at the input layer. A multitude of procedures and combinations are used to build the networks. MobileNetV2 and NasNetMobile are 2 DL models for smart phones. MobileNetv2's design has 155 layers as well as 164 connections [33, 34]. Separable convolutions are employed in mobile design, are utilized in MobileNetv2. NasNetMobile's mobile edition is divided into twelve sections. NasNet is a flexible CNN composed of fundamental construction components enhanced using recurrent neural networks [35]. A cell is made up of only a few actions that are frequently duplicated due to the network's required size. The layer Global Average Pooling [36] was used, which significantly minimises forwarding error prediction failure.

SqueezeNet is a small network designed to provide a more compacted alternative to AlexNet [37]. It has nearly 50 times less parameters than AlexNet yet performs three times quicker. SqueezeNet's core ideas are as follows:

- Approach one is to use 1×1 filters instead of 3×3 filters.
- Approach second: decrease the input channels to 3×3 filters.
- Approach three: reduce the network late in the process such that the convolution layers have huge activation maps.

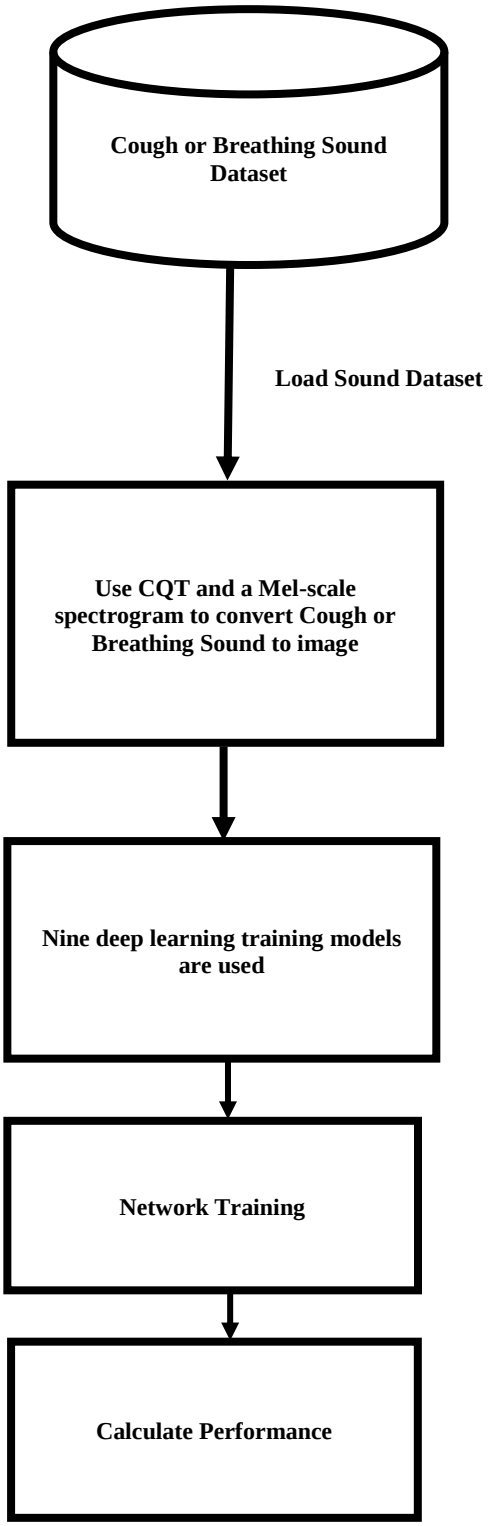
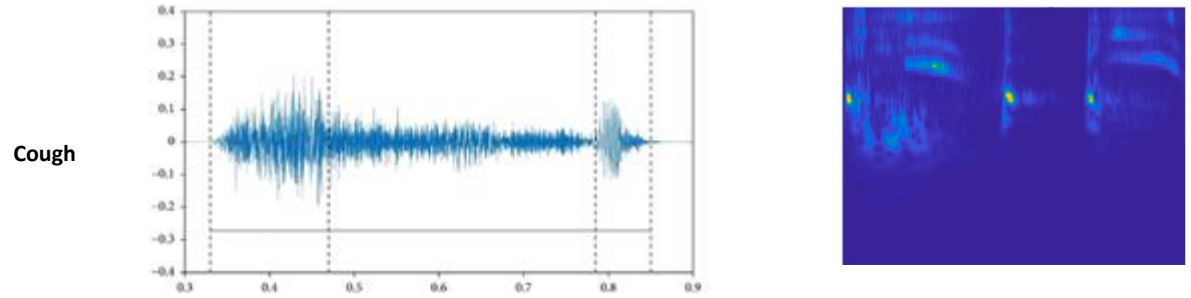
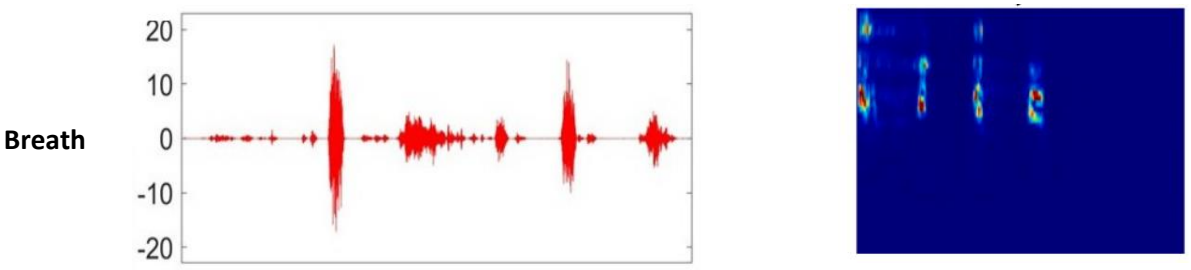


Fig. 2. The Presented COVID-19 Coughing and Breathing DL Classification Model.

The SqueezeNet design contains 15 layers, with 5 distinct layers and 2 convolution layers.



(a) Healthy



(b) COVID-19

Fig. 3. A Mel-Scale Spectrogram and Constant-Q Transform of an Electrocardiogram of Covid-19 and Healthy Cough and Breath Sounds.

The Residual Network is a well-known deep learning model (ResNet). The development of these Residual blocks lessened the difficulty of training very deep networks, and the ResNet model is founded on them. ResNet provides several models, like 18/34/50/101/152. ResNet18 has 18 convolutional layers and a 33 filter. The ResNet-34 design entailed placing shortcut links onto a plain network in order to convert it into its residual network counterpart. In this scenario, the plain network was impacted by VGG neural networks (VGG-16, VGG-19) with a 33 filter in the convolutional networks. The ResNet-34 design contains 34 convolution layers. Regardless of the fact that the Resnet50 architecture is centered on the preceding generation, it differs in one significant way. Large Residual Networks, like as ResNet101 (101 layers) as well as ResNet152 (152 layers), are constructed utilising extra 3-layer blocks. Also when network depth is raised, the 152-layer ResNet had significantly decreased complexity.

GoogLeNet is constructed on numerous extremely small convolutions to significantly reduce the amount of parameters. The GoogLeNet architecture contains 22 layers, although the parameters have been reduced from 60 million (AlexNet) to 4 million. GoogLeNet has nine inception modules to investigate clustering and network inside a network. During the inception modules, the module range is computed, and the entirely connected layers are deleted. Pooling parameters in the inception modules decreases the number of parameters. In furthermore, a shadow network and also an auxiliary classifier were used to enhance the findings [38].

A CQT is used to transform a time-domain signal to a frequency-domain signal, that is then evaluated with many resolutions. The use of a Mel-scale spectrogram to display signal characteristics, in addition to its ability to distinguish biometrically, distinguishes this paper. In light of this, the signal processing system maintains its morphological difficulties. This implies that ML based on basic classifiers may be unsuccessful in identifying complex signals. We sent an image through CNN's DL, which showed to be the most effective in detecting visual morphology. The DL model output has not been equal. As a result, the current study intended to create the most representative DL models for image categorization (GoogleNet, ResNet18/34/50/100/101, MobileNetv2, SqueezeNet, and NasNetmobile).

V. Experimental Results

The provided DL model is performed in transfer mode using the suggested basic training setup (batch norm epsilon= e^{-3}, weight decay= $\boldsymbol{e^{-3}}$, and batch norm decay= 0.5, and dropout= 0.5). The batch size= $\mathbf{8}$, as well as the learning rate = $\mathbf{0.02}$, which was lowered till it reached $\boldsymbol{e^{-5}}$ automatically. The Deep Learning models are tested for 20 hours on a DELL PC with a 2.4 GHz Intel Core (TM) i7-M520 CPU, MATLAB R2016 64-bit, and 16 GB RAM running Windows 10 as well as tensorflow's Deep Neural Network library (CuDNN).

The dataset was divided into three parts: 80% for training, 10% for validation, and 10% for testing. We used both labelled and assessment data in our investigation. Validation accuracy is a classification score that is used to assess the learning technique as it proceeds. The size of the dataset determines the split ratio. To ensure the maximum level of model efficiency, an appropriate balance between training and testing must be attained. Furthermore, there is no immediate

reaction to the process or parameter pushes one over the brink. The results of each DL transfer model are shown in **Table I**, with an initial learning rate of 0.02 and 22 epochs. The batch size was set to eight, and early ending was permitted if there was no change in accuracy. It was revealed that by using more samples, the model output improved [39]. Stochastic Gradient Descent with Momentum (SGDM) [40] was the optimizer technique used in this study to improve detector performance. To prevent over-fitting issues with the Deep Learning net, we adopted the dropout approach [41]. As indicated in eq. (3), the teaching criteria were the loss function $L(x,t)$, which is defined as the total of binary plus box loss functions. Also, Eq. (5) and (6) are used to calculate the regression loss L_{re}:

$$L(x,t) = L_{cl}(x_c) + \delta[b > 0]L_{re}(z,z^*) \quad (3)$$

Where, $(\boldsymbol{z_a}, \boldsymbol{z_b}, \boldsymbol{z_w}, \boldsymbol{z_h})$indicates the bounding boxes of $\boldsymbol{z}$ and $\boldsymbol{z}^*$, $\boldsymbol{w}$ as well as $\boldsymbol{h}$ signify the box's width and height respectively, and $\boldsymbol{x_c}$ denotes the predicted score class $\boldsymbol{c}$. Non-background boxes at zero are defined by $\boldsymbol{\delta}[\boldsymbol{b} > \boldsymbol{0}]$. The bounding box as well as the classification loss $\boldsymbol{L_{cl}}$, are involved in the regression loss, as seen in eq. (4).

$$L_{cl}(x_{c^*}) = -\log(x_{c^*}) \quad (4)$$

$$L_{re}(k,k^*) = \sum_{i\in(a,b,w,h)} V_{LI}\,(k_i - k_i^*) \quad (5)$$

Where,

$$V_{LI}(p) = \begin{cases} 0.5p^2, if\ |p| < zero \\ |p| - 0.5,\ otherwise \end{cases} \quad (6)$$

A. Examination of Performance

Testing can yield a positive result, demonstrating the Deep Learning models' dependability. The confusion matrix is a statistical performance calculation approach used in study. Among the six statistical metrics are accuracy, sensitivity, specificity, precision, the F1 score, and the Matthews Correlation Coefficient (MCC). **Fig. 4** and **Fig. 5** show the confusion matrices for the two categories (COVID-19 as well as Healthy). Eq. 7 was used to get as close to the truth as feasible

$$accuracy = [N_{TP} + N_{TN}]/[(N_{TP} + N_{FP}) + (N_{TN} + N_{FN})] \quad (7)$$

Where, N_{TP}, N_{FN}, N_{TN}, and N_{FP} are No. of correctly labeled, mislabeled, clearly labelled instances of the remaining classes and incorrectly labelled instances of the remaining classes respectively. The efficiency of the five ResNet models (ResNet 18/34/50/100/101) is shown in Fig. 4(a, b, c, d, e), and the overall accuracy is 98.9%, 91.4%, 93.1%, 92.9%, and 90.1%. The confusion matrix of the test for the GoogleNet model is given in Fig. 5(a), as well as the accuracy rate is 89.9%. Fig. 5(b, c, d) depicts the performance of MobileNetv2, NasNetMobile, and SqueezeNet, with accuracy rate of 89.2%, 88.9%, and 86.9%, respectively. Because of the tiny dataset, Resnet18 obtained the maximum accuracy. The accuracy of DL models' predictions was quantitatively tested. Both sensitivity and accuracy are widely utilized classification efficiency metrics. Eq. 8 and 9 are applied to calculate Sensitivity and Precision. Fig. 6 depicts the sensitivity and specificity of the nine Deep Learning models.

ResNet101 has a sensitivity of 95.6% when it relates to distinguishing COVID-19 persons' breathing sounds. ResNet18 does indeed have 98.1% specificity, meaning that it can detect people who do not have COVID-19. Precision, F1 score, and MCC are calculated using Eq. 10, 11, and 12. **Fig. 6** depicts the accuracy, F1 score, as well as MCC for the nine Deep Learning models. ResNet18 does have the highest accuracy of 96.4%, indicating that it generates more relevant findings than the other models. In **Fig. 6**, the DL model's efficiency is assessed by a test with a high F1 score of 95.9% for ResNet18. Finally, the MCC demonstrates that the more statistically reliable rate performed well in all four categories of the uncertainty matrix. ResNet18 has the highest MCC of 91.8 percent.

$$Sensitivity = N_{TP}/(N_{TP} + N_{FN}) \quad (8)$$

$$Specifcity = \frac{N_{TN}}{(N_{TN}+N_{FP})} \quad (9)$$

$$Precision = N_{TP}/(N_{TP} + N_{FP}) \quad (10)$$

$$F1\ score = 2 \times N_{TP}/(2 \times N_{TP} + N_{FP} + N_{FN}) \quad (11)$$

$$MCC = \frac{N_{TN}\times N_{TP} - N_{FP}\times N_{FN}}{\left(\sqrt{(N_{TP}+N_{FP})\times(N_{TP}+N_{FN})\times(N_{TN}+N_{FP})\times(N_{TN}+N_{FN})}\right)} \quad (12)$$

B. Examination of Performance Discussions and Comparative Analyses

In Fig. 6, the outcomes of the suggested method for applying deep learning DL models in the breathing dataset implementing CQT as well as Mel-scale spectrogram images of COVID-19 illness and healthy are shown. Fig. 7 shows how our proposed approach can effectively recognise data. The present study's innovations include the employment of CQT and a Mel-scale spectrogram with deep learning models to characterise signal characteristics and biometric identification capabilities. The core of related research focuses on classifying breathing and coughing signals through ML. Table II analyses the performance of several techniques in terms of accuracy. The authors in [23, 24] employed a small dataset that included the actual COVID-19 coughing sound sample in a comparable investigation. Much of the prior study has been on distinguishing between coughing and non-coughing tones. We noticed this when analysing the effectiveness of Deep Learning transfers methods in detecting COVID-19 cough sounds using the SGDM, when cough signals occur often, the efficiency of all Deep Learning techniques improves significantly. While having the greatest performance, our detection model's efficiency is just 98.9% relying on the SGDM optimizer, the learning data correctness and the effort to analyse the labelled data. Every inaccuracy in data recording that evaded our notice, on the other hand, is most likely to influence the reported outcomes. Table III demonstrates the accuracy of the proposed suggested model when implemented to the Coswara dataset. The state-of-the-art models are labelled in the first left column. According to Table III, the proposed technique achieves high accuracy compared to the other models. The findings of this study, as well as those of another study mentioned in the related works section, indicate that specific latent properties of coughing sounds may well be successfully exploited for DL identification of a variety of respiratory issues. As it

differentiates between normal and COVID-19 coughing, the coughing can be used as a preliminary diagnostic technique. We study the use of a Mel-scale spectrogram of tone as a return to Deep learning to see if the model is greater than effective at identifying medical images to tone. ResNet as well as GoogleNet were proved to have great accuracy in this work despite being recognised as deep variants of DL transfer models. For mobile versions, NasnetMobile as well as mobilenetv2 provide great precision. For assessment, the tests are done on a separate dataset that consists of audio wave files. Resnet18 was much more effective than GoogleNet, while resnet34/50/100/101 was much more effective than GoogleNet. NasnetMobile outperformed Mobilenetv2m and Squeezenet in terms of accuracy. The case is used in the experiments to assess the existing classification model's efficiency and consistency. According to the results, the resnet18 model has the greatest classification accuracy on cough as well as breath signals from the confirmed COVID-19 dataset. The DL classification outperforms conventional CNN classifications in matching coughing and breathing sounds of COVID-19 sufferers. As an outcome, it could really aid in diagnosis by alleviating clinicians of the stress connected from the first sound of the COVID-19 cough as well as breath.

TABLE I. DEEP LEARNING MODELS SETUP

Deep Learning models	**LAYERS**	**Batch Size**	**Epoch**	**Learning rate**	**Optimizer**
Googlenet	20	8	22	0.02	SGDM
ResNet18	18				
ResNet34	34				
ResNet50	50				
ResNet100	100				
ResNet101	101				
MobileNetv2	53				
NasNetMobile	cells				
SqueezeNet	15				

TABLE II. IN TERMS OF ACCURACY, A COMPARISON OF SEVERAL METHODOLOGIES IS MADE

Reference	**LAYERS**	**Dataset**	**Result**
[23]	CNN	1427	80.7%
[24]	CNN	871	70.5%
[25]	CNN	317	92.6%
[29]	CNN	1457	94.9%
Current study	**Deep Learning transfer model**	**1850**	**98.9%**

TABLE III. IN TERMS OF ACCURACY, DISPLAYS OUTCOMES FOR VARIOUS MODELS FOR CLASSIFICATION ON COSWARA DATASET

Reference	**LAYERS**	**Result**
[23]	CNN	85.6%
[24]	CNN	72.9%
[25]	CNN	87.1%
[29]	CNN	91.4%
Current study	**Deep Learning transfer model**	**98.7%**

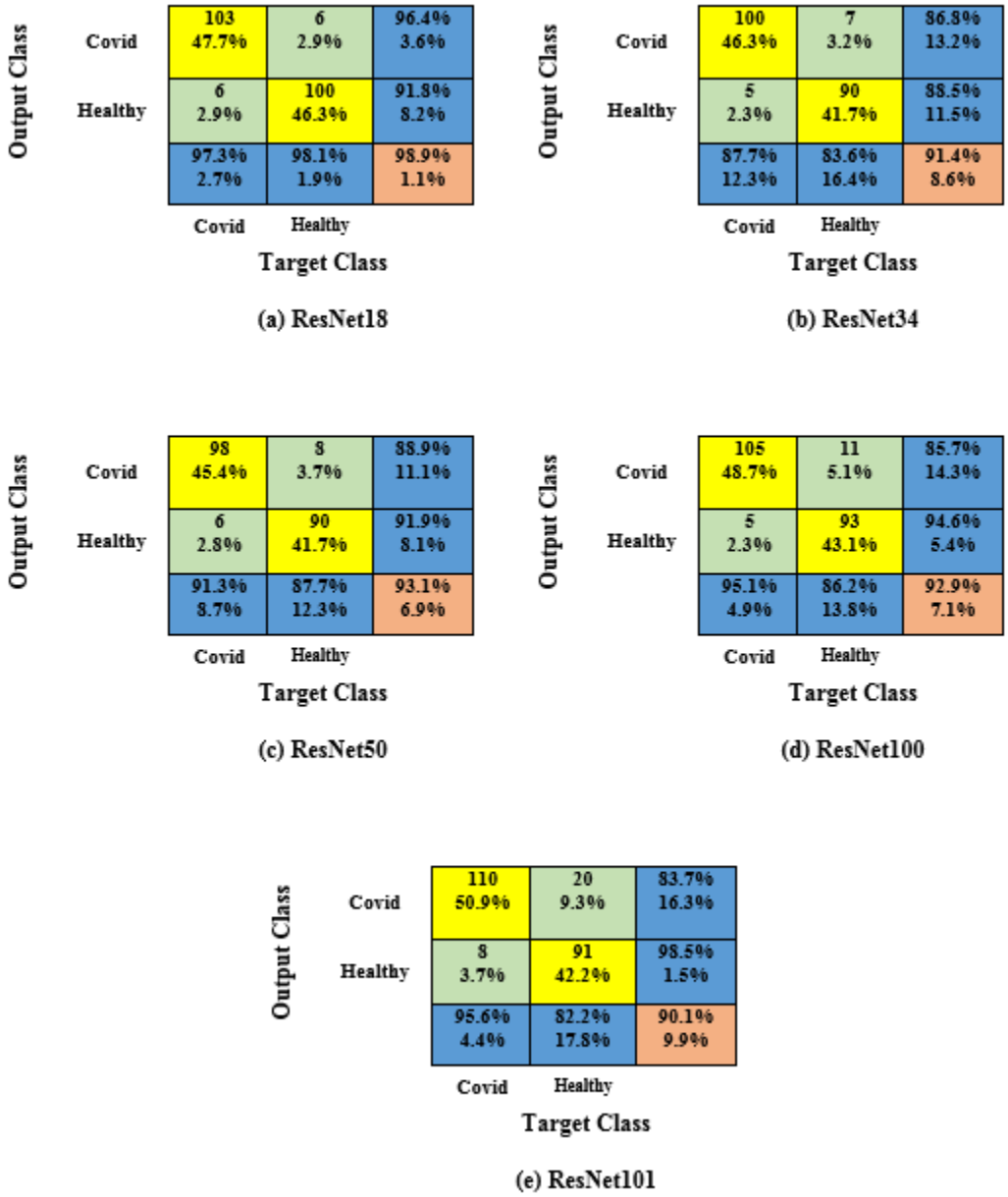


Fig. 4. Shows Confusion Matrix of ResNet18/34/50/100/100.

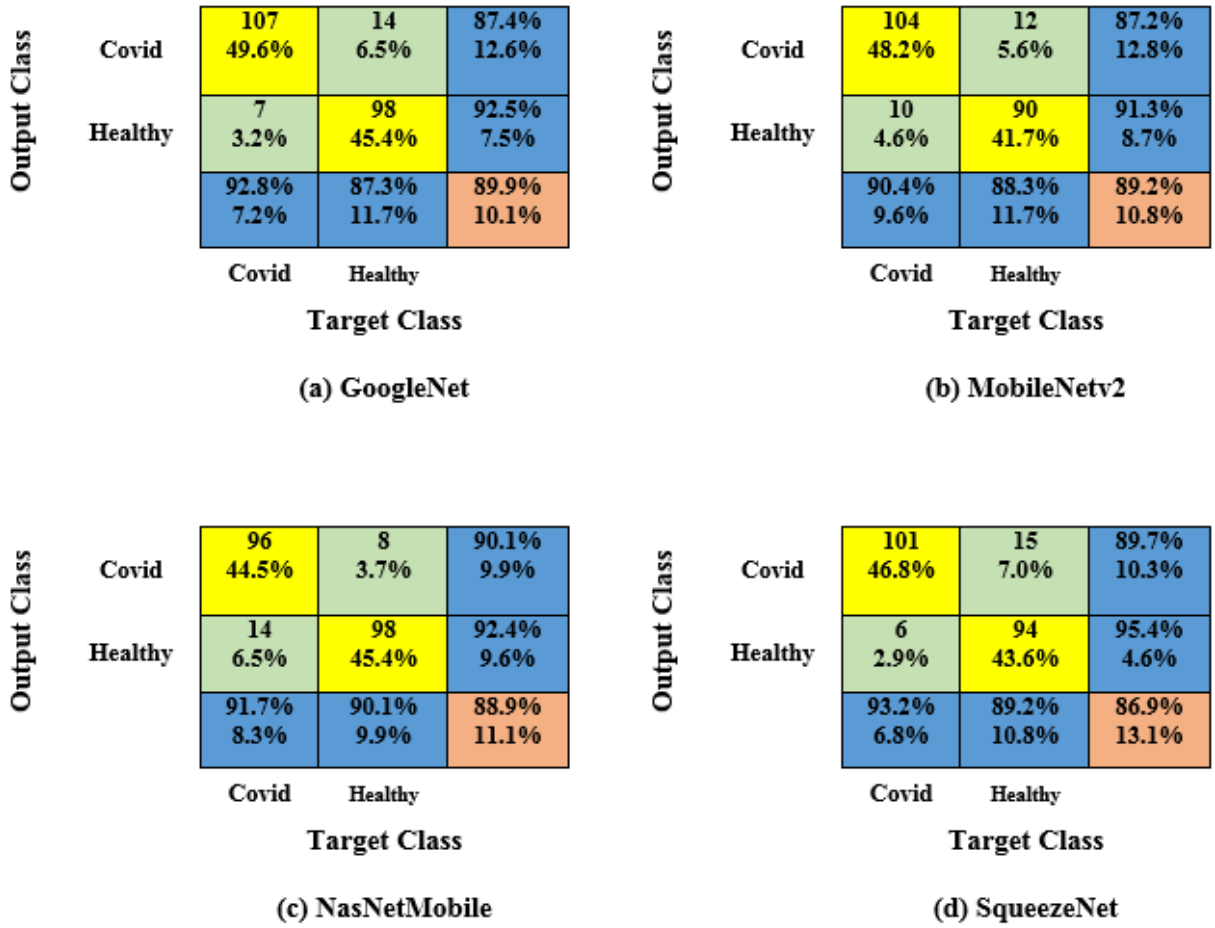


Fig. 5. Shows Confusion Matrix of GoogleNet, MobileNetv2, NasNetMobile, and SqueezeNet.

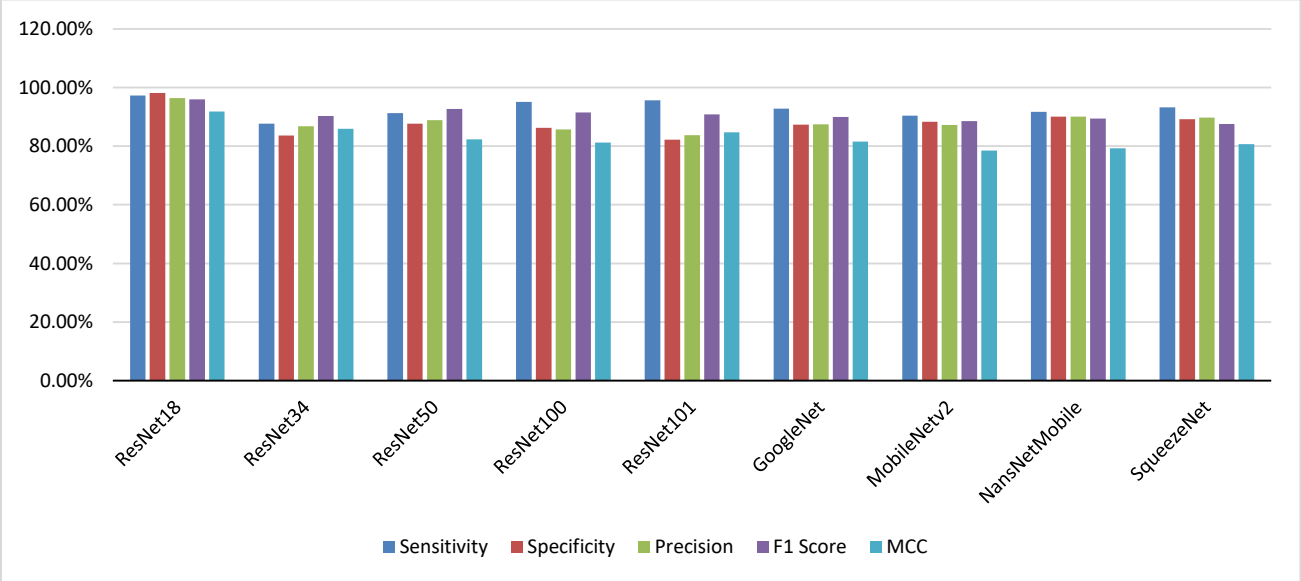


Fig. 6. Shows Sensitivity, Specificity, Precision, F1 Score, and MCC for All Deep Learning Models.

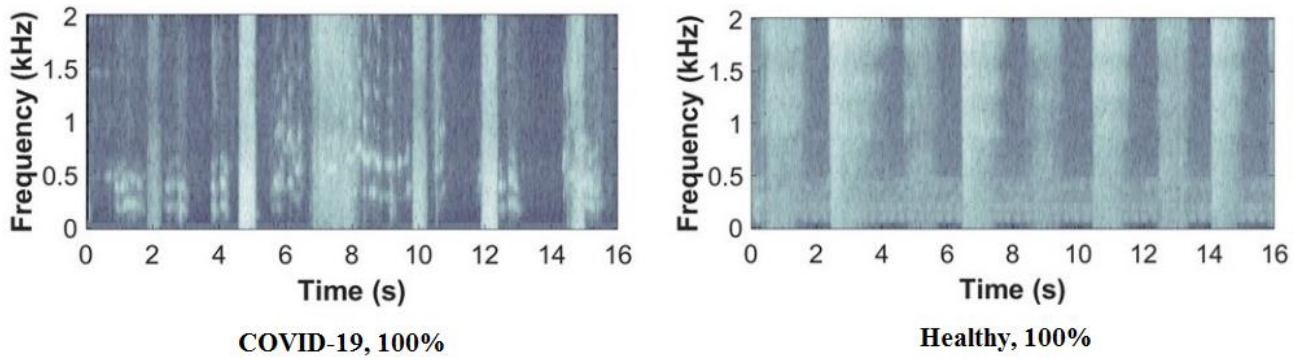


Fig. 7. Shows Samples of Shallow Breathing Tones with their CQT and Mel- Scale Spectrograms.

VI. CONCLUSION

The current study created innovative DL models for breath and cough sound classification that focus on sound and might aid in COVID-19 transmission controls. The proposed model combines two key components. The initial method was using a CQT as well as a Mel-scale spectrogram to transform sound waves into images. The second component is the construction of universal features as well as extra classification utilising deep transfer models (GoogleNet, ResNet18, ResNet34, ResNet50, ResNet100, ResNet101, MobileNetv2, SqueezeNet and NasNetmobile). Around 1,600 people from all over the world (1185 men and 415 women) supplied data to the collection (mostly the Indian population). With the better classification performance employing coughing sound CQT and a Mel-spectrogram image, the current proposal outperformed the other nine CNN networks. For symptomatic patients, the accuracy, sensitivity, and specificity were 98.9%, 97.3%, and 98.1%, respectively. The Resnet18 network is the most dependable for symptomatic patients who use coughing as well as breathing tones. When applied to the Coswara dataset, we discovered that the suggested model's accuracy (98.7%) outperforms the state-of-the-art models (85.6%, 72.9%, 87.1%, and 91.4%) based on the SGDM optimizer. The suggested study's findings contribute to key suggestions for future ML and DL research. Our results can be comparable to a scalogram, another common type of time-frequency representation. Given its high overall accuracy, the suggested study will require more replication before it may be used in other healthcare applications. This work opens the door for the use of DL in COVID-19 diagnosis by proving that it is a rapid, time-efficient, and low-tech solution that does not violate social separation criteria in pandemics like COVID-19.

ACKNOWLEDGMENT

This research received no external funding.

REFERENCES

[1] WHO Coronavirus, COVID-19, Dec. 30, 2021. [Online]. Available: https://covid19.who.int.

[2] M. Loey, F. Smarandache, and N.E.M. Khalifa, "Within the lack of chest COVID-19 X-ray dataset: a novel detection model based on GAN and deep transfer learning," Symmetry, vol. 12, no. 4, Apr. 2020.

[3] Mohammad-Rahimi H, Nadimi M, Ghalyanchi-Langeroudi A, Taheri M, and Ghafouri-Fard S, " Application of machine learning in diagnosis of COVID-19 through X-ray and CT images: a scoping review," Frontiers in cardiovascular medicine, Vol. 8, Mar. 2021.

[4] Ozkaya U, Ozturk Ş, Barstugan M,"Coronavirus (COVID-19) classification using deep features fusion and ranking technique. In: Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach," Springer, p. 281–295, 2020.

[5] Wang X, Deng X, Fu Q, Zhou Q, Feng J, and Ma H,"A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT," IEEE transactions on medical imaging, vol. 39, no. 8, pp. 2615–2625, 2020.

[6] Majid A, Khan MA, Nam Y, Tariq U, Roy S, and Mostafa RR," COVID19 classification using CT images via ensembles of deep learning models," Computers, Materials and Continua, p. 319–337, 2021, [Online]. Available: https:// doi.org/10.32604/cmc.2021.016816.

[7] Khan MA, Kadry S, Zhang YD, AkramT, SharifM, and Rehman A," Prediction of COVID-19-pneumonia based on selected deep features and one class kernel extreme learning machine," Computers & Electri- cal Engineering, vol. 90, 2021.

[8] AkramT, Attique M, Gul S, Shahzad A, Altaf M, and Naqvi SSR," A novel framework for rapid diagnosis of COVID-19 on computed tomography scans," Pattern analysis and applications, pp. 1–14, 2021.

[9] ZhaoW, JiangW, and Qiu X," Deep learning for COVID-19 detection based on CT images," Scientific Reports, vol.11, no. 1, pp.1-12, 2021.

[10] Khan MA, AlhaisoniM, Tariq U, Hussain N, Majid A, and Damas ˇevičius R,"COVID-19 Case Recognition from Chest CT Images by Deep Learning, Entropy-Controlled Firefly Optimization, and Parallel

Feature Fusion," Sensors, vol. 21, no. 21, 2021.

[11] Shui-HuaW, Khan MA, Govindaraj V, Fernandes SL, Zhu Z, and Yu-Dong Z,"Deep rank-based average pooling network for COVID-19 recognition," Computers,Materials, & Continua, p. 2797–2813, 2022.

[12] Zhang YD, Khan MA, Zhu Z,Wang SH," Pseudo zernike moment and deep stacked sparse autoencoder for COVID-19 diagnosis," Cmc-Computers Materials & Continua, p. 3145–3162, 2021.

[13] Kaushik H, Singh D, Tiwari S, Kaur M, Jeong CW, and Nam Y," Screening of COVID-19 patients using deep learning and IoT framework," Cmc-Computers Materials & Continua., pp. 3459–3475, 2021.

[14] V. Bhateja, A. Taquee, and D.K. Sharma,"Pre-processing and classifcation of cough sounds in noisy environment using SVM," in: 2019 4th International Conference on Information Systems and Computer Networks (ISCON), pp. 822–826, Nov. 2019.

[15] W. Gao, W. Bao, and X. Zhou,"Analysis of cough detection index based on decision tree and support vector machine," J. Combin. Optim., vol. 37, no. 1, pp. 375–384, Jan. 2019.

[16] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton,"Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012.

[17] S. Liu, and W. Deng,"Very deep convolutional neural network based image classification using small training sample size," in: 3rd IAPR Asian Conference On Pattern Recognition, pp. 730–734, 2015.

[18] C. Szegedy," Going deeper with convolutions," in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1–9, 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun,"Deep residual learning for image recognition," in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[20] R.X.A. Pramono, S.A. Imtiaz, E. Rodriguez and Villegas,"A cough-based algorithm for automatic diagnosis of pertussis," PLoS One, vol. 11, no. 9, Sep. 2016.

[21] N. Rochmawati," Covid symptom severity using decision tree," in: Third International Conference On Vocational Education And Electrical Engineering (ICVEE), pp. 1 5, Oct. 2020.

[22] M. Soli´ nski, M. Łepek, and Ł. Kołtowski," Automatic cough detection based on airflow signals for portable spirometry system,"Informatics in Medicine Unlocked, Jan. 2020.

[23] B.W. Schuller, H. Coppock, and A. Gaskell,"Detecting COVID-19 from Breathing and Coughing Sounds Using Deep Neural Networks," arXiv:2012.14553 [cs, eess], Dec. 2020, Accessed: Jan. 16, 2021, [Online]. Available: http://arxiv.org/abs/2012 .14553.

[24] V. Bansal, G. Pahwa, and N. Kannan,"Cough Classifcation for COVID-19 based on audio mfcc features using Convolutional Neural Networks," in: IEEE International Conference On Computing, Power And Communication Technologies (GUCON), pp. 604–608Oct. 2020.

[25] Imran,"AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," Informatics in Medicine Unlocked ,2020.

[26] J.-M. Liu, M. You, Z. Wang, G.-Z. Li, X. Xu, and Z. Qiu,"Cough event classifcation by pretrained deep neural network," BMC Med. Inf. Decis. Making, vol. 15, no. 4, Nov. 2015.

[27] M. Aly, and N. Alotaibi, "A novel deep learning model to detect COVID-19 based on wavelet features extracted from Mel-scale spectrogram of patients' cough and breathing sounds, " Informatics in Medicine Unlocked, Vol. 32, 2022.

[28] Y. Amrulloh, U. Abeyratne, V. Swarnkar, and R. Triasih, "Cough Sound Analysis for Pneumonia and Asthma Classifcation in Pediatric Population,"in Modelling And Simulation 2015 6th International Conference On Intelligent Systems, pp. 127–131, Feb. 2015.

[29] M. Loey and S. Mirjalili,"COVID-19 cough sound symptoms classifcation from scalogram image representation using deep learning models," Computers in Biology and Medicine, vol. 139, Nov. 2021.

[30] Sharma N, Krishnan P, Kumar R, Ramoji S, Chetupalli SR, and Ghosh PK," Coswara–A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," arXiv, 2020.

[31] Indian institute of science,"Project Coswara," [Online]. Available: https://coswara.iisc.ac.in/team.

[32] Paul S Addison,"Wavelet transforms and the ECG: a review," Physiological Measurement, Nov. 2005.

[33] H. Yasar, and M. Ceylan,"A novel comparative study for detection of Covid-19 on CT lung images using texture analysis, machine learning, and deep learning methods," Multimed Tool Appl., Oct. 2020.

[34] I.D. Apostolopoulos, S.I. Aznaouridis, and M.A. Tzani,"Extracting possibly representative COVID-19 biomarkers from X-ray images with deep learning approach and image data related to pulmonary diseases,"J. Med. Biol. Eng., vol. 40, no. 3, pp. 462–469, Jun. 2020.

[35] B. Zoph, V. Vasudevan, J. Shlens, and Q.V. Le,"Learning Transferable Architectures for Scalable Image Recognition, " arXiv:1707.07012 [cs, stat], Apr. 2018, Accessed: Jan. 17, 2021, [Online]. Available: http://arxiv.org/abs/1707.07012.

[36] C. Szegedy, S. Ioffe, V. Vanhoucke, and A.A. Alemi,"Inception-v4, inception-ResNet and the impact of residual connections on learning, " in: Proceedings Of the Thirty-First AAAI Conference On Artifcial Intelligence, San Francisco, California, USA, , pp. 4278 -4284, Feb. 2017, pp. 4278 -4284.

[37] V.K. Pothos, D. Kastaniotis, I. Theodorakopoulos and N. Fragoulis," A fast, embedded implementation of a Convolutional Neural Network for Image Recognition," Technical Report, Aug. 2016, [Online]. Available: https://www.researchgate.net/publication/306003694_A_fast_embedded_implementation_of_a_Convolutional_Neural_Network_for_Image_Recognition.

[38] Nur Ateqah, Nur Hidayah, Zaidah Ibrahim, and Nur Nabilah," Celebrity Face Recognition using Deep Learning," Indonesian Journal of Electrical Engineering and Computer Science, vol. 12, no. 2, pp. 476-481, Nov. 2018.

[39] Y. Xu, and R. Goodacre,"On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," J. Anal. Test, vol. 2, no. 3, pp. 249-262, Jul. 2018.

[40] Sutskever, J. Martens, G. Dahl, and G. Hinton," On the importance of initialization and momentum in deep learning," in: Proceedings of the 30th International Conference on International Conference on Machine Learning, vol. 28, 2013.

[41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov,"Dropout: a simple way to prevent neural networks from overftting," J. Mach. Learn. Res., vol. 15, no. 56, pp. 129-1958, 2014.

[42] K. Khoria, et.al, "On significance of constant-Q transform for pop noise detection," Computer Speech & Language, Vol. 77, 2023.

MFCC and Texture Descriptors based Stuttering Dysfluencies Classification using Extreme Learning Machine

Roohum Jegan, R. Jayagowri
Department of Electronics and Communication Engineering
BMS College of Engineering
Bangalore, India

***Abstract*—Stuttering is a type of speech disorder which results in disrupted flow of speech in the form of unintentional repetitions and prolongation of sounds. Stuttering classification is important for speech pathology treatment and speech therapy techniques which decreases speech disfluency to some extent. In this article, a method for prolongation and repetition classification is presented based on Mel-frequency cepstral coefficients (MFCC) and texture descriptors. Initially, MFCC and filter bank energy (FBE) matrix are computed. Gray level co-occurrence matrix (GLCM) and Gray level run length matrix (GLRLM) textural features are extracted from these matrices. Laplacian score-based feature selection approach is employed to choose relevant features. Finally, extreme learning machine (ELM) is utilized to classify the speech audio event as repetition or prolongation. The algorithm is evaluated using UCLASS database and has achieved improved performance with classification accuracy of 96.36%.**

Keywords*—Voice disorder; Mel-frequency cepstral coefficients; gray level co-occurrence matrix; GLRLM; Laplacian score; extreme learning machine

I. Introduction

Speech forms a major part in day-to-day communication and is used by humans to express their emotions and to exchange their ideas. Thus, speech helps in the efficient communication of ideas that determines how a person thinks and feels. Speech is a special gift to the mankind because animals and other species cannot speak [1]. Stuttering is classified as one of the speech disorders and is identified by reiteration of utterances, phonics, phrases, or terms; elongation of sounds during utterance; and interventions in speech called as blocks [2]. Even though there is no complete cure for stuttering at present, there are numerous speech pathology approaches that may aid to decrease speech disfluency to certain extent. To judge the performance of the stutterers before and after the treatment, stuttering assessment is needed. Generally, Speech Language Pathologist (SLP) manually enumerates and categorize eventuality of disfluencies such as prolongations and repetitions in a stammered speech. But, this type of evaluation is unpredictable, uncertain, intuitive, cumbersome and erroneous. Hence it would be worthy if the stuttering assessment can be carried out automatically enabling the SLP to spend more time with the stutterer in treatment session.

This article presents a new statistical feature based on MFCC and FBE matrix to enhance stuttering event classification using UCLASS database. Prolongation and repetition event are discriminated using GLCM and GLRLM features extracted from MFCC and FBE matrix and ELM classification. Laplacian score- based feature selection algorithm is employed to discard irrelevant features resulting in improvement in the classification rate. The proposed feature extraction approach improves the prolongation and repetition classification accuracy to a greater extent. Moreover, best features are selected using Laplacian score-based feature selection algorithm, thereby, minimizing the computational complexity.

The rest of the paper is organized as follows: In Section II, the past solutions for the predetermined problem via different algorithms, classification and feature extraction techniques are presented. In Section III, the proposed method for speech dysfluencies has been discussed with brief description of each method in separate subsections. In Section IV, the simulation results of the work are discussed, and Section V presents the conclusion of the work with the future scope.

II. Literature Review

This section presents different objective approaches proposed for stuttering event classification based on various features, datasets and classifiers. In [3], automatic detection of syllable repetitions is presented using correlation of 1/3 octave spectra. The correlation features are used to identify repeated syllables with similar spectral components. Acoustic and pitch related descriptors including MFCCs, formants, tonality (pitch), zero crossing rate (ZCR) and energy are employed to classify repetitions and prolongations using Artificial Neural Networks (ANN) in [4]. The accuracy obtained using the ANN based classification was 87.39%. Line spectral frequency (LSF) representation features are extracted and classified using three different classifiers: MLP, RNN and RBF resulting 98-100% detection rate in [5].

LP-Hilbert transform based MFCC (LH-MFCC) based feature extraction method is presented to classify three different dysfluencies using Gaussian Mixture Model (GMM) classifier [6]. These features efficiently capture temporal and spectral parameters of utterances resulting in 94.98% accuracy rate. To enhance classification accuracy, a decision fusion

technique is introduced, based on combination of different acoustical features like ZCR, speech envelope (ENV) for classifying filled pause (FP) and elongation (ELO) in Malay language [7]. Stuttered speech repetition detection algorithm based on MFCC and dynamic time warping (DTW) with accuracy of 83-90% is proposed in [8], [9].

Various stuttering events are classified in [10] using i-vector based KNN and LDA classification resulting in 80- 85% classification accuracy. In [11], similarity matrix image is computed using MFCC, PLP and filter bank energy feature sets. Dysfluent regions are detected using threshold based morphological image processing having an average classification accuracy of 82.5%. SVM based dysfluency classification method using a GMM supervector is introduced in [12], [13] with +96.10% accuracy. Repetition and prolongation classification using MFCC, LPC and perceptual linear predictive (PLP) and k-NN and SVM classifier is proposed in [14], [15] having classification rate of 96%.

A method using SVM classifier and fusion of prosodic (pitch and energy) and cepstral (MFCC) features is presented for stuttered speech classification with 97.80% accuracy in [16]. A deep belief network architecture is developed based on MFCC and LPCC features to classify stuttering speech signal having an accuracy of 85% in [17]. Computational intelligence approach based on ANN and SVM is developed to classify dysfluencies in stuttered speech signal in [18] with 85% accuracy rate. An objective methodology for dysfluency detection using six-level wavelet packet transform decomposition and features employing entropy features is presented in [19], [20]. Performance of the algorithm is evaluated using three distinct classifiers including k-NN, LDA and SVM classifiers resulting in classification accuracy of 96.67%. MFCC and LPCC based stuttered event classification approach is proposed in [21] using k-NN and LDA classifiers with 94% classification rate. Prolongation and repetition in stuttered speech classification technique using LPCC features and k-NN/LDA classifier is presented with 89.77% in [22].

In another study, [23], [24], same authors presented stuttered event detection approach using LPC, LPCC and WLPCC features with 97.78% classification accuracy. But the test segments taken were very small and it was observed that accuracy decreased for bigger test segments. MLP network architecture is presented in [25] to detect stop consonant repetitions with accuracy of 76.67%. This study presents a new approach based on MFCC and FBE matrix representation and feature extraction using GLCM and GLRLM descriptors. Convolutional Method (CNN) was used to classify different languages in [42]. The classification accuracy obtained is 97.86%.

The existing literature feature extraction approaches are primarily based on time-domain features extracted from the speech sample. The stuttering classification rate is limited between 87% and 94%. Additionally, the feature selection techniques are less explored in the existing literature hence limiting the classification accuracy. This article presents a new feature extraction algorithm based on MFCC and FBE matrix statistical features. The proposed feature extraction approach enhances the prolongation and repetition accuracy. Moreover, important features are selected using Laplacian score-based feature selection algorithm, thereby, reducing the computational complexity.

III. Proposed Method for Speech Dysfluencies Classification

This article presents prolongation and repetition classification using MFCC/FBE, two different types of texture descriptors and ELM classifier. The proposed scheme is presented in this section. MFCC and filter bank energy computation is explained next along with its importance in the stuttering classification process. GLCM and GLRLM texture features are investigated and discussed in detail. Laplacian score-based feature selection is employed in this study along with its brief introduction. Finally, extreme learning machine classifier that is employed in this work and the merits are discussed.

A. Architecture of MFCC and FBE based Dysfluencies Classification

The framework of the stuttering classification scheme is depicted in Fig. 1. It is the architecture used for stuttering classification using MFCC and texture descriptors. The sample voice is analyzed before taking it as an input from a person. At various stages, the input speech signal is manipulated and undergoes operations of Pre-processing, converting into frames, filtering and Windowing, and complementing with the uttered word. This speech algorithm has two major stages: training and testing stages and the process is shown in Fig. 2.

After pre-processing the input speech sample, MFCC and FBE matrix is obtained. GLCM and GLRLM descriptors are extracted from these two matrices. In order to reduce feature vector dimensionality, Laplacian score-based filter type feature selection is used. Finally, ELM is trained using the training database. In our experiments, 70% of the speech specimens are used for training stage and remaining 30% samples for testing.

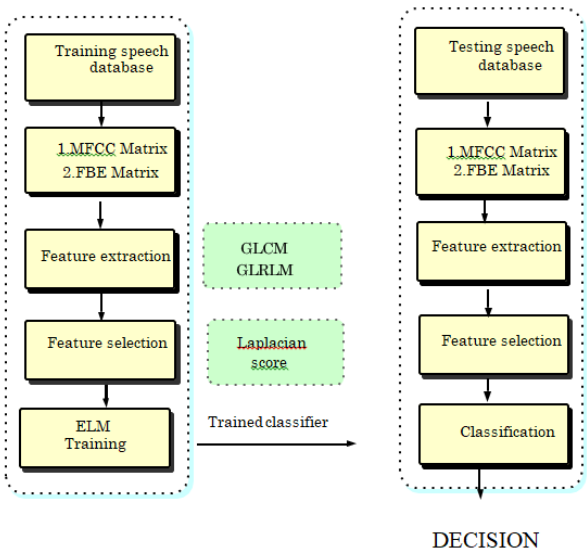


Fig. 1. Architecture of Stuttering Classification using MFCC and Texture Descriptor.

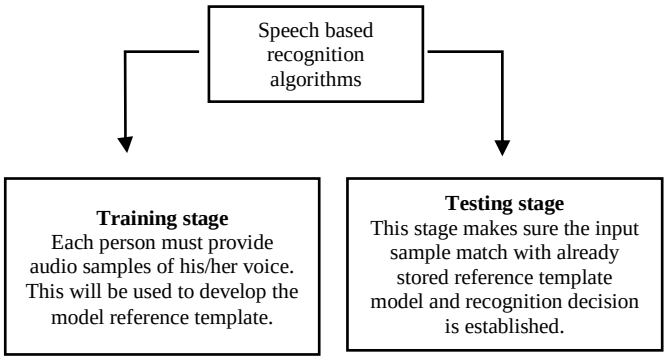


Fig. 2. Training and Testing of Voice Recognition Algorithms.

B. Mel-Frequency Cepstral Coefficients (MFCCs)

MFCC represents set of Discrete Cosine Transform decorrelated variables that are evaluated using transmutation of the energies (output from filter) that are compressed logarithmically. These are obtained from a sharply spaced triangular filter bank that precedes the Discrete Fourier Transformed audio signal. The extracted features represent parametric characterization of audio signals that plays important role to enhance the performance of the recognition approach. MFCCs is widely and commonly adopted feature extraction algorithm in variety of audio/speech/music processing algorithms [26]-[30].

MFCC describes short time cepstral features and uses Mel scale with linear separation below 1000 Hz and logarithmic spacing above 1 kHz. The value of Mel for any frequency f (Hz) is computed as:

$$\text{Mel}(f) = 2595 \times \log10(1 + f/700) \quad (1)$$

where M is the quantity of triangular filters, L represents the total Mel scale coefficients and Ek is energy of the filter bank (log) output. Filter bank approach characterizes the speech samples efficiently. A set of triangular band-pass filter is designed, and nonlinear Mel-frequency scale is employed considering human perceptual capabilities with specific frequency spacing. Intensity from each band is computed by multiplying Mel filter bank and magnitude spectrum of speech signal. We observed that, filter bank energy spectrum varies based on the input speech sample (prolongation and repetition). This dissimilarity is exploited in this study for stuttering event classification. The overall architecture and process of generating Mel frequency cepstral coefficients is shown in Fig. 3 [6], [7]:

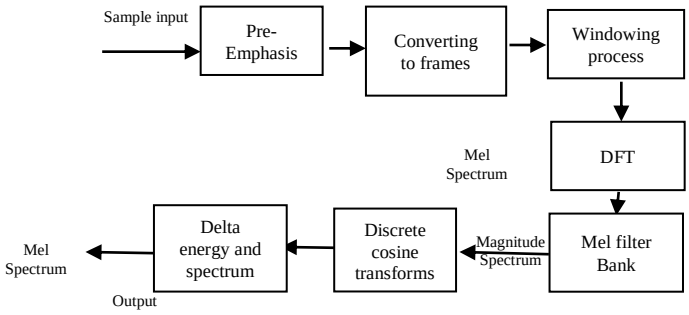


Fig. 3. MFCC Flow.

MFCC constitutes of seven major stages. Every stage has its own mathematical processes and functions described in the following steps:

Step 1: Pre–emphasis: This stage helps in escalating the signal energy at greater frequency range by allowing the signal to pass through a filter that prioritizes higher frequencies.

$$Y[x] = M[x] - aM[x - 1 \quad (2)$$

$$Y[x] = M[x] - 0.96M[x - 1] \quad (3)$$

Let us consider 'a' to have a value 0.96, which means it is assumed that there is 96% chance of a sample to regenerate from the previous sample.

Step 2: Converting to frames: Framing refers to segmentation of the voice samples into small framework with distance between 20 to 40 ms. The speech samples are derived from the analog to digital convertor. The input audio signal is segmented into N sample frames. Nearby frames are segregated by Y (Y < M). Y = 100 and M = 256 are the most typical values used.

Taking into account the succeeding step in feature extraction stage, it amalgamates all the immediate frequency bands. If the window is given as H(x); where x = count of samples contained in each frame, B(x) = signal output, A(x) = signal input, H(x) = window. The, the output of hamming window is:

$$B(x) = A(x) \times H(x) \quad (4)$$

$$h(x) = 0.54 - 0.46 \cos \Sigma 2\pi x;\ 0 \le x \le X - 1 \quad (5)$$

Step 4: Fast Fourier Transform: It transforms each framework of X samples that are in time domain to corresponding frequency domain. FT transforms the convolution of pulse in glottis tt[n] and impulse response I[n] of the vocal tract present in the time domain which is shown in equation stated below:

$$B(x) = \text{FFT}[I(t) * A(t) = I(x) * A(x)] \quad (6)$$

If A(x), I(x) and B(x) are the FT of A(t), I(t) and B(t) respectively.

Step 5: Filter Bank Conversion: The FFT range has high frequency and is broad and the audio signal is non-linear. Fig. 4 shown above describes a set of triangular filters to enumerate the weighted sum of all filter spectral samples so that the output is made to approach the Mel scale. Every filter has triangular magnitude frequency response with unit value at the center frequency and it gradually reduces linearly to zero at the Centre frequency of adjoining filters [7], [8]. Output of each filter is the filtered sum of its spectral components. Equation stated below is then used to calculate the Mel for any frequency f (HZ) as:

$$\text{Mel}(f) = 2595 \times \log10(1 + f/700) \quad (7)$$

Step 6: DCT: Log Mel spectrum is converted into time domain using Discrete Cosine Transform and this result in the formation of MFCC. MFC coefficients are also called the acoustic vectors. Hence, each input speech sample is converted into a chain of acoustic vector.

Step 7: Delta Energy and Delta Spectrum: The audio speech signal and the frameworks vary in accordance to the formant slope at its changeovers. Hence, features that relates to the variations in cepstral parameters over time have to be included. 13 delta parameters that include 12 cepstral features and one energy feature, and 39 double delta features are included. The energy E of a signal 'a' in a window frame from time duration t11 to time sample t12, is given by the following equation:

$$E = A^2t \quad (8)$$

Each of the thirteen delta variables constitutes the variation happening between frames corresponding to the energy parameter. On the other hand, 39 double delta features depict the variations among frames in the corresponding delta features as,

$$r(t) = [s(t + 1) - s(t - 1)]/2 \quad (9)$$

In short, the MFCC computation comprises of the framing stage where the input pre-processed speech sample is divided into several frames with overlap. After framing the signal, hamming window attenuates the framed signal to null at the beginning and end of the frame. The windowed signal is converted to frequency domain by applying Fast Fourier transform (FFT). The FFT spectrum is passed through a set of triangular band-pass filter to obtain the logarithm energy spectrum.

The placement of these filters is based of Mel frequency scale, which is proportional to logarithm of linear frequency scale, reflecting human perceptual capabilities. In the last step, discrete cosine transform (DCT) is applied on logarithm energy to extract L Mel scale cepstral coefficients using the energy compaction property and is obtained as,

$$C(n) = \Sigma\, Ek \times \cos(n \times (k - 0.5) \times \pi/40 \quad (10)$$

Fig. 4 and 5 shows filter bank energy spectrum for prolongation and repetition samples from UCLASS database respectively. It is clearly seen that for prolongation samples, (Fig. 4) the filter bank energy of the frames for prolongation utterances are equal. Whereas, in case of repetition (Fig. 5), filter bank frame energies are distributed more evenly as compared to prolongation frame energies. Moreover, central coefficient energy distribution is higher in Fig. 5 compared to Fig. 4. These differences are exploited for the classification in this article.

C. Gray Level Co-occurrence Matrix (GLCM)

Feature extraction technique is mainly used to make simpler the number of features that can accurately describe a large set of data. While analyzing complex data, large number of variables involves more difficulties. Huge number of variables traditionally requires more memory and computational power. Else it requires a classification algorithm which can fit the entire training sample but that result in poor generalization of new samples. Feature extraction refers to methods used for establishing fusions of the features while still describing the data without compromising on the accuracy. It is mainly used in applications that describes and retains the texture kinesthetic or visual attributes of a surface.

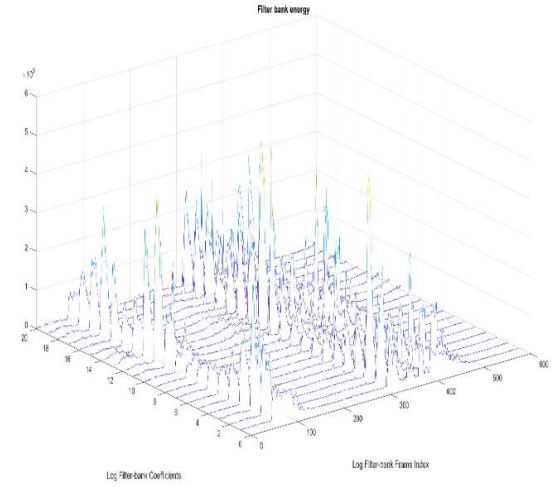


Fig. 4. Filter Bank Energy Spectrum for Prolongation Samples.

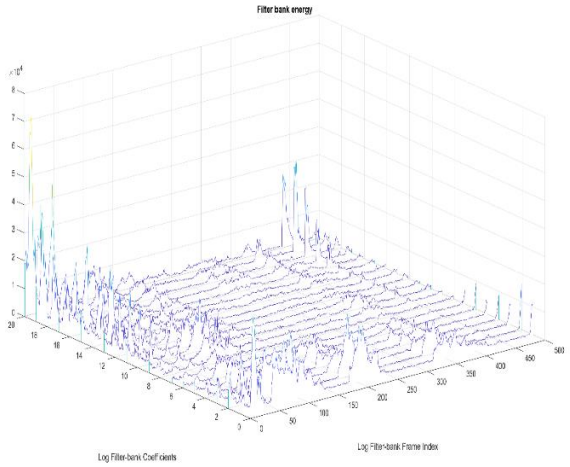


Fig. 5. Filter Bank Energies for Repetition Samples.

Texture analysis helps to find a distinctive way of illustrating the underlying/hidden characteristics of texture and personify them in an effortless and distinct form. This helps in robust, accurate classification and segmentation of samples. Although textural representation contributes a major role in image study and pattern recognition, only a few architectures implement the idea of onboard textural feature extraction method. Gray level co-occurrence matrix helps to obtain qualitative/statistical texture features. Hence GLCM is a numerical analysis method used for observing the texture that contemplates the spatial correlation between pixels [31]. GLCM is one of the most widely used texture descriptor to compute statistical features of the image based on gray level intensities and employed in different image processing applications like image segmentation, image retrieval, image classification and object recognition as discussed in [41].

The main advantage of the co-occurrence matrix computation is that, while considering the relation between two pixels at any instant of time, pixel pairs that coincide can be topologically matched in various inclinations in accordance to the distance and spatial-based angular relationships. This apparently exhibits the combination of grey levels of image

matrix and their positions. Matrix relationships are defined by changing the directions with different angles and displacement vectors. In this paper, twenty features are extracted from the MFCC and FBE matrix [32].

The count of rows and number of columns in a GLCM is the count of gray levels, H, present in the image. The matrix component A (a, b δp, δq) is the frequentness at which two pixels, divided by an interval (δp, δq), transpire within a given vicinity, one with potency 'a' and the other with potency 'b'. The component A (a, b δp, δq) has got the values of second order statistical probability for corresponding variations among the gray levels 'a' and 'b' at a specific displacement length l and at a specific angle (θ).

GLC Matrices are very delicate to the dimension of the texture sample on which they are computed because of their large dimensionality. Hence, the reduction in the number of gray levels is of utmost importance. In most of the cases, a reference pixel and its quick neighbor is contemplated. Usage of larger offset is feasible if the window is large enough. The top most cell at the left will contain the frequency of occurrence of combination 0,0. That means, it will contain the information about the total number of times a neighbor pixel having 0 gray level gets placed to the right of reference pixel with 0 gray level, within the image area.

D. Gray Level Run Length Matrix (GLRLM)

This matrix is another popularly used higher order texture descriptor useful in feature extraction. It is a textural representation model that helps in extracting the spatial plane features of each pixel relative to the higher order statistics [33]. A 2- dimensional feature matrix is obtained at the end of the process, exploiting the spatial variations because of the prolongation and repetitions. GLRL matrix is not just limited to 00 direction. It can also be used in other directions with θ = 450, θ = 900 and θ = 1350.

GLRL matrix gives us few textural parameters that can be extracted from it. It is observed that five texture features can be extracted from this GLRL matrix, namely: Shot Runs Emphasis (SRE), Long Runs Emphasis (LRE), Non-uniformity Gray Level (GLN), Non-uniformity Run Length (RLN), and Percentage of Run (RP). Later on, two more features called the Low Gray Level Run Emphasis (LGRE) and High Gray Level Run Emphasis (HGRE) were found to be extracted from this matrix. This parameter makes use of sequential gray level of pixels and then discriminates the texture that has equal values for Shot Runs Emphasis and Long Runs Emphasis with minor variations in the gray level distribution.

After those four more features were found to be extracted from the matrix, namely: Low Short Run Gray-Level Emphasis (SRLGE), High Short Run Gray Level Emphasis (SRHGE), Low Long Run Gray Level Emphasis (LRLGE), and High Long Run Gray Level Emphasis (LRHGE). Same intensity adjacent pixels in certain direction is termed as run length. Each element in the GLRLM characterizes the total gray level occurrences in the given direction. For a single MFCC matrix, it is possible to compute many different run- length matrices f (i, j θ) one for each chosen direction θ. Thus, given a direction, for each acceptable gray measure value, this matrix measures the total run times. GLRLM is parameterized by three different pixel features: intensity, length and direction of a run from a reference pixel. Total of 11 features per direction are extracted from the MFCC and FBE matrix [34]-[36].

E. Laplacian Score based Feature Selection

Feature selection plays important role as the preprocessing step in machine learning to select optimum features from the large input feature set. Feature selection techniques can be classified into: (a) filter and (2) wrapper techniques [37]. Filter methods are independent of the learning algorithm and faster, whereas, wrapper approaches produce higher accuracy and it needs learning algorithm. In this article, Laplacian score- based filter approach is used for feature selection [38]. As the name suggests, for every parameter, its Laplacian score is evaluated and calculated separately to reveal its locality preserving power.

Laplacian score approach is based on the reflection that two data points are probably related to the same point if they are near to each other. Generally, in all the learning tasks like classification, the local geometric structure is more important than the global structure of the given feature space. Hence, a nearest neighbor graph is designed to construct the local structure, and Laplacian score aspires those specific parameters that obey this graph model.

Laplacian score (LS) is based on the concepts of 'Laplacian Eigenmaps' and 'Locality Preserving Projection' and is used to identify importance of individual features. Locality preserving power is computed using this LS for each feature and the features are inferred to be similar if they produce very low LS. Based on the graph, structure is defined using the nearest neighbor and the geometric structure of the descriptor is evaluated. As LS is a ranking filter approach for feature selection, a threshold T is used to select number of features for classification.

F. Extreme Learning Machine (ELM) Classifier

Extreme learning machine algorithm is a contemporary state-of-the art machine learning algorithm with sole-hidden layer feed-forward neural network (SLFNs). ELM is fast; it has better generalization performance and enhances the training speed by assigning the weights randomly. ELM requires only two parameters: (1) hidden layer neural units and (2) their transfer function [39] and [43]. ELM algorithm is used in data classification and regression applications. The optimal values must be chosen for ELM training parameters to enhance the accuracy. However, while designing the classifier using ELM, the number of hidden nodes to be used for handling different problems remains a trial and error [40].

A major drawback of ELM is that the classifying borderline for the learning features of this algorithm may not be an adequate one. This is because the learning features of hidden nodes are arbitrarily allocated while they remain uninterrupted in the training stage [17]. Hence, few features might be miscategorized by the algorithm, mainly for those samples that are close to the classifying border line. Another observation made is that, in many cases, this algorithm might need additional hidden neurons compared to the already available traditional tuning-based algorithms [18]. Few researchers have

proposed that the above-mentioned shortcomings of ELM can be overcome by introducing several variants of ELM, such as incremental ELM [9], pruning ELM [12], error-minimized ELM [19], dual-step ELM [20], sequential online ELM [21], evolutionary ELM [18], voting-based ELM [17], ordinal and fully complex ELM [23], and balanced (symmetric) ELM.

IV. Simulation Results

Proposed stuttering event classification algorithm is evaluated using speech samples from UCLASS database [3]. The database includes 43 speakers recording generating 107 audio samples.

In this article, 39 speech samples are selected for classification similar to the settings used in [21]. During MFCC and FBE computation, the analysis frame duration is set to 25 ms with overlap of 10 ms. The pre-emphasis coefficient is set to 0.97 with 20 filter bank channels and 12 cepstral coefficient extractions. Lower frequency limit is set to 300 Hz, whereas 3700 Hz is the high frequency limit.

The GLCM features are extracted using one direction only, as during the experimentation we found that one direction is sufficient to generate satisfactory classification rate. GLCM descriptors are extracted from both MFCC and FBE matrix representation, generating 20-D feature vector for each. In addition to GLCM, GLRLM textural features are also extracted from these two matrices, thereby generating 24-D feature vector for MFCC and FBE matrix individually. Out of the total speech samples, 70% are used to train the ELM and enduring 30% are employed for testing. ELM is implemented using 300 neurons, which is set experimentally with sigmoidal transfer function. Finally, each stuttering speech sample is represented using 84-D feature vector.

Table I shows prolongation and repetition classification accuracy for individual (GLCM and GLRLM separately) and combined feature sets (GLCM+GLRLM). From Table I, it is observed that, GLCM has poor discrimination capability with classification accuracy of only 79.84%. Compared to GLCM features, GLRLM descriptors are more powerful during the classification. Finally, as expected, combined feature set (GLCM+GLRLM) resulted in highest accuracy of 92.64%. It is also evident that, feature fusion enhances the classification rate notably in the stuttering event discrimination. In order to demonstrate the effect of Laplacian score feature selection, additional experiments are performed. Table II depicts prolongation and repetition classification accuracy for individual (GLCM and GLRLM separately) and combined feature set using Laplacian score feature selection approach. Significant improvement in the classification accuracy can be observed (see Table II) by applying the feature selection technique. Combined (GLCM+GLRLM) feature set accuracy obtained was 96.36% using only 25 features. Thus, Laplacian score approach not only enhances the classification rate but decreases the number of features also (60% decrease in total number of features).

As this approach is ranking based approach, threshold T is used to select number of important features from the large input feature set. Fig. 6 shows the classification accuracy obtained using different threshold values. Highest accuracy was obtained at threshold T = 0.2, we choose final feature set with this threshold (resulting in final 25-dimensional relevant features only). As started above, GLCM descriptors can be evaluated using four different directions. Table III illustrates GLCM detection accuracy using different directions employing LS feature selection and without LS feature selection. As evident, feature selection improves the detection rate. It is also worth mentioning that, combining all four directions enhances the detection rate of the proposed technique. The present work utilizes ELM classifier with 300 hidden neurons. Tables IV and V shows the number of hidden neurons and corresponding obtained accuracy using GLCM and GLRLM features, respectively. It was observed that the highest classification rate is achieved using 300 neurons.

TABLE I. Prolongation and Repetition Classification Accuracy for Individual and Combined Feature Set

Features	Classification Accuracy (%)
GLCM	79.84
GLRLM	84.6
Combined	92.64

TABLE II. Prolongation and Repetition Classification Accuracy for Individual and Combined Feature Set using LS feature Selection

Features	Classification Accuracy (%)
GLCM	81.74
GLRLM	87.24
Combined	96.36

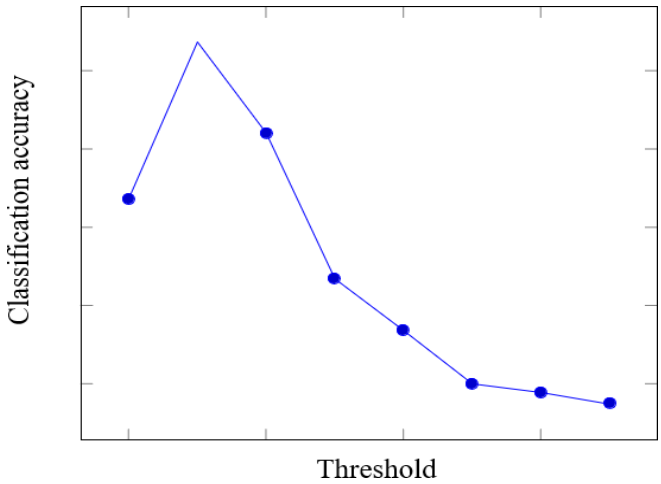


Fig. 6. Classification Accuracy vs Threshold Values using Laplacian Score-based Feature Selection for Stuttering Event Classification.

TABLE III. GLCM Detection Accuracy using Different Directions

Direction	Accuracy without FS	Direction	Accuracy with FS
00	79.84	00	81.74
00 + 450	78.12	00 +450	81.24
00+450+900	78.05	00+450+900	80.71
00+450+900+1350	77.69	00+450+900+1350	79.93

TABLE IV. NUMBER OF HIDDEN NEURONS AND ACCURACY USING GLCM FEATURES

Number of hidden neurons	Classification Accuracy (%)
100	75.29
150	76.35
200	75.85
250	78.16
300	79.84
350	78.97

TABLE V. NUMBER OF HIDDEN NEURONS AND ACCURACY USING GLRLM FEATURES

Number of hidden neurons	Classification Accuracy (%)
100	79.20
150	80.93
200	81.21
250	83.79
300	84.60
350	83.64

V. DISCUSSION

The proposed MFCC and FBE based textural feature approach is compared with existing stuttering event classification algorithms. Table VI depicts comparison of proposed method with already existing state-of- the art methods using different features and classification accuracy rates. It can be seen that the proposed technique performs better compared to all the traditional algorithms.

TABLE VI. COMPARISON OF PROPOSED METHOD WITH EXISTING METHODS USING DIFFERENT FEATURES AND CLASSIFICATION RATE

Method	Features	Classification Accuracy (%)
[4]	Acoustic and pitch	87.39
[6]	LH-MFCC	94.98
[8]	MFCC-DTW	90
[9]	MFCC-DTW	89
[14]	MFCC, LPC	95
[15]	MFCC, LPC, PLP	96
[21]	MFCC, LPCC	94
Proposed	MFCC, FBE	96.36

The prime objective of this article is to present a new statistical feature approach based on MFCC and FBE matrix to enhance stuttering event classification using UCLASS database. Prolongation and repetition event are distinguished using GLCM and GLRLM features extracted from MFCC and FBE matrix and Extreme learning machine classification. Laplacian score-based feature selection algorithm is employed to remove irrelevant features resulting in improvement in the classification accuracy rate. Experiments show that GLRLM outperforms GLCM descriptors during the classification stage. On selecting the best feature set of 25 features (T = 0.2), highest accuracy of 96.36% is obtained. And it can be observed from the results and tables that the performance of the proposed algorithm is better compared to other existing methods. Besides, this article also emphasizes the use of feature selection technique to reduce the computational complexity of the algorithm.

VI. CONCLUSION

This article presents stuttering event classification approach based on MFCC and FBE using UCLASS database. Prolongation and repetition event are discriminated using GLCM and GLRLM features extracted from MFCC and FBE matrix and ELM classification. Laplacian score-based feature selection algorithm is employed to discard irrelevant features resulting in improvement in the classification rate. Experimental results show that, GLRLM outperforms GLCM descriptors during the classification stage. After selecting best feature set of 25 features (T = 0.2), highest accuracy of 96.36% is achieved. In future works, experiments can be performed with large speech samples with different feature extraction approaches and classifiers to improve the classification rate further.

REFERENCES

[1] "Statistics on Voice, Speech, and Language NIDCD" https://www.nidcd.nih.gov/health/statistics/statistics-voice-speech-and-language, Accessed: 2020-12-10.

[2] W. Suszynski, W. Kuniszyk-Jzkowiak, E. Smoka, and M. Dzienkowski, "Speech disfluency detection with the correlative method," Annales UMCS Informatica, vol. 3, no. 1, pp. 131 – 138, 2005.

[3] P. Howell, S. Davis, and J. Bartrip, "The University College London archive of stuttered speech (UCLASS)," Journal of Speech, Language, and Hearing Research, vol. 52, no. 2, pp. 556 – 568, 2009.

[4] P. S. Savin, P. B. Ramteke, and S. G. Koolagudi, "Recognition of repetition and prolongation in stuttered speech using ANN," in Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, A. Nagar, D. P. Mohapatra, and N. Chaki, Eds. Springer India, 2016, pp. 65–71.

[5] N. K. A. M. Rashid, S. A. Alim, N. N. W. N. Hashim, and W. Sediono, "Receiver operating characteristics measure for the recognition of stuttering dysfluencies using line spectral frequencies," International Islamic University Malaysia Engineering Journal, vol. 18, no. 1, pp. 193–200, 2017.

[6] P. Mahesha and D. S. Vinod, "LP-Hillbert transform based MFCC for effective discrimination of stuttering dysfluencies", in 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), March 2017, pp. 2561–2565.

[7] R. Hamzah, N. Jamil, and R. Roslan, "Development of acoustical feature-based classifier using decision fusion technique for malay language dis- fluencies classification", Indonesian Journal of Electrical Engineering and Computer Science, vol. 8, no. 1, pp. 262–267, 2017.

[8] P. B. Ramteke, S. G. Koolagudi, and F. Afroz, "Repetition detection in stuttered speech," in Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, A. Nagar, D. P. Mohapatra, and N. Chaki, Eds. New Delhi: Springer India, 2016, pp. 611–617.

[9] P. Yeh, S. Yang, C. Yang, and M. Shieh, "Automatic recognition of repetitions in stuttered speech: Using end-point detection and dynamic time warping," Procedia - Social and Behavioral Sciences, vol. 193,

[10] p. 356, 2015, 10th Oxford Dysfluency Conference, ODC 2014, 17 - 20 July, 2014, Oxford, United Kingdom.

[11] A. G. Samah, A. Sherif, S. Mahmoud, and G. Nivin, "Classification of stuttering events using I-Vector," Egyptian Journal of Language Engineering, vol. 4, no. 1, pp. 11–18, 2017.

[12] I. Esmaili, N. J. Dabanloo, and M. Vali, "Automatic classification of speech dysfluencies in continuous speech based on similarity measures

and morphological image processing tools", Biomedical Signal Processing and Control, vol. 23, pp. 104 – 114, 2016.

[13] M. P. and V. D. S., "Support vector machine-based stuttering dysfluency classification using GMM supervectors," Int. J. Grid Util. Comput., vol. 6, no. 3/4, pp. 143–149, 2015.

[14] H. M., C. L. Sin, A. O. Chia, and Y. Sazali, "Gaussian mixture model-based classification of stuttering dysfluencies," Journal of Intelligent Systems, vol. 25, no. 3, pp. 387–399, 2015.

[15] K. Singh and A. K. Awasthi, "Comparison of speech parameterization techniques for the classification of speech disfluencies," Turkish Journal of Electrical Engineering and Computer Science, vol. 21, pp. 1983 – 1994, 2014.

[16] M. P. and D. S. Vinod, "Classification of speech dysfluencies using speech parameterization techniques and multiclass SVM," in Quality, Reliability, Security and Robustness in Heterogeneous Networks, Eds. Springer Berlin Heidelberg, 2013, pp. 298–308.

[17] J. L. C., P. Srikanta, and I. Nikhil, "Combining cepstral and prosodic features for classification of disfluencies in stuttered speech," in Intelligent Computing, Communication and Devices, Eds. Springer India, 2015, pp. 623–633.

[18] O. Stacey, M. Ricard, and R. Frank, "Automatic dysfluency detection in dysarthric speech using deep belief networks," in Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies. Association for Computational Linguistics, 2015, pp. 60–64.

[19] J. Palfy, "Analysis of dysfluencies by computational intelligence," Information Sciences and Technologies-Bulletin of the ACM Slovakia, vol. 6, no. 2, pp. 45–58, 2014.

[20] M. Hariharan, V. Vijean, C. Y. Fook, and S. Yaacob, "Speech stuttering assessment using sample entropy and least square support vector machine," in 2012 IEEE 8th International Colloquium on Signal Processing and its Applications, March 2012, pp. 240–245.

[21] M. Hariharan, C. Fook, R. Sindhu, A. H. Adom, and S. Yaacob, "Objective evaluation of speech dysfluencies using wavelet packet transform with sample entropy," Digital Signal Processing, vol. 23, no. 3, pp. 952– 959, 2013.

[22] L. S. Chee, O. C. Ai, M. Hariharan, and S. Yaacob, "Automatic detection of prolongations and repetitions using LPCC," in 2009 International Conference for Technical Postgraduates (TECHPOS), Dec 2009, pp. 1– 4.

[23] H. M., C. L. Sin, A. O. Chia, and Y. Sazali, "Classification of speech dysfluencies using LPC based parameterization techniques," J. Med. Syst., vol. 36, no. 3, pp. 1821–1830, 2012.

[24] L. S. Chee, O. C. Ai, M. Hariharan, and S. Yaacob, "Automatic detection of prolongations and repetitions using LPCC," in 2009 International Conference for Technical Postgraduates (TECHPOS), Dec 2009, pp. 1– 4.

[25] wietlicka Izabela, K.-J. Wiesawa, and S. Elbieta, "The application of Kohonen and multilayer perceptron networks in the speech non-fluency analysis," Archives of Acoustics, vol. 31, 01 2006.

[26] A. Benba, A. Jilbab, A. Hammouch, and S. Sandabad, "Voiceprint's analysis using MFCC and SVM for detecting patients with Parkinson's disease, "in 2015 International Conference on Electrical and Information Technologies (ICEIT), March 2015, pp. 300–304.

[27] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. OShaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," Speech Communication, vol. 55, no. 2, pp. 237 – 251, 2013.

[28] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-variance multitaper MFCC features: A case study in robust speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 7, pp. 1990–2001, Sept 2012.

[29] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1085–1095, May 2012.

[30] M. Sahidullah and G. Saha, "A novel windowing technique for efficient computation of MFCC for speaker recognition," IEEE Signal Processing Letters, vol. 20, no. 2, pp. 149–152, Feb 2013.

[31] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610–621, Nov 1973.

[32] L. Soh and C. Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," IEEE Transactions on Geoscience and Remote Sensing, vol. 37, no. 2, pp. 780–795, March 1999.

[33] M. Galloway, "Texture analysis using gray level run lengths," Computer Graphics and Image Processing, vol. 4, no. 2, pp. 172–179, June 1975.

[34] X. Tang, "Texture information in run-length matrices," IEEE Transactions on Image Processing, vol. 7, no. 11, pp. 1602–1609, Nov 1998.

[35] B. V. Dasarathy and E. B. Holder, "Image characterizations based on joint gray level run length distributions," Pattern Recognition Letters, vol. 12, no. 8, pp. 497–502, August 1991.

[36] A. Chu, C. M. Sehgal, and J. F. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," Pattern Recognition Letters, vol. 11, no. 6, pp. 415–419, June 1990.

[37] L. Zhu, L. Miao, and D. Zhang, "Iterative Laplacian score for feature selection," in Pattern Recognition, C.-L. Liu, C. Zhang, and L. Wang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 80–87.

[38] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in Advances in Neural Information Processing Systems 18, Y. Weiss, B. Schlkopf, and J. C. Platt, Eds. MIT Press, 2006, pp. 507–514.

[39] A. Bequ and S. Lessmann, "Extreme learning machines for credit scoring: An empirical evaluation," Expert Systems with Applications, vol. 86, pp. 42 – 53, 2017.

[40] Z. Zhou, Y. Song, Z. Zhu, and D. Yang, "Scene categorization based on compact SPM and ensemble of extreme learning machines," Optik vol. 6, no. 2, pp. 45–58, 2014.

[41] O. C. Ai, M. Hariharan, S. Yaacob, and L. S. Chee, "Classification of speech dysfluencies with MFCC and LPCC features," Expert Systems with Applications, vol. 39, no. 2, pp. 2157 – 2165, 2012.

[42] Gajanan K. Birajdar, Vijay H. Mankar "Passive Image Manipulation Detection Using Wavelet Transform and Support Vector Machine Classifier", Proceedings of International Conference on ICT for Sustainable Development pp 447-455, 2016.

[43] Gajanan K. Birajdar & Mukesh D. Patil, "Speech and music classification using spectrogram based statistical descriptors and extreme learning machine", Multimedia Tools and Applications vVolume78, pages15141–15168 (2019).

Innovation Management Model as a Source of Business Competitiveness for Industrial SMEs

Rafael Roosell Paez Advincula[1], Celso Gonzales Chavesta[2], Lilian Ocares-Cunyarachi[3]
Ingeniería Industrial[1], Ingeniería Estadística e Informática[2], Ingeniería de Sistemas e Informática[3]
Image Processing Research Laboratory (INTI-Lab), Universidad de Ciencias y Humanidades (UCH)
Lima, Perú

***Abstract*—One of the main problems of small companies is not knowing how to properly manage investments, resources, strategies, tools and responsibilities to be more competitive, the research objective is the development of management and innovation processes required by the new company for its permanence in the market and decision making to be carried out every day; small companies face the competitiveness of new products that emerge. Today there are a great variety of products, which are globally interconnected, therefore it is required to implement a structural equation model for its management, so that the company continuously improves and optimizes the available resources, therefore as a result is the management, focused on greater effectiveness of resources; so it is necessary that small businesses need to manage a model of investments, resources, tools and responsibilities to obtain support from the competitiveness of the market; which would allow it to be oriented to a sustainable development and be one step ahead of the competition.**

Keywords*—Competitiveness; continuous improvement; management; optimization; strategies

I. Introduction

Currently innovation has taken great importance in the small business sector to be sustainable in the market must have an analysis of the variables that determine the operation of a process in the execution of actions of the main factors that allow to achieve the goals, objectives and effectively complete the assignment tasks for which they were programmed, to achieve this should be considered to control complex processes with the help of different components for continuous improvement and have business success in a competitive environment [1].

In order to be competitive in the market, productivity must be increased and the quality of the products must be improved to achieve this, the implementation of technology and innovation that allows greater speed and optimal production is applied companies that work with various tools, devices and process control equipment must maintain the controls in operation for it is necessary to apply reliable and efficient tools [2].

For the role of innovation, it is necessary for the company to apply knowledge management, in order to obtain a higher level of commitment to achieve added value and be more competitive to achieve this, efficient and appropriate tools are needed to allow immediate solutions to problems, process control helps us to identify a system of continuous improvement for decision making in the company and optimize resources and goods, so as to ensure the effectiveness and efficiency in the process centers, opting for a mechanism for sustainable development, which was designed as a factor of organizational competitiveness [3].

A new work scheme allows to improve the processes and activities in a way to acquire a sustainable development from the point of view of the industry this has widely developed models of continuous improvement that allow to define the processes dedicated to the engineering components and where the most used technique is applied to break down the vision of the business in the strategic objectives, the establishment of different control panels allows to apply a model of development of techniques to manage the system, mainly of the processes of programming models and perform the analysis of the levels of innovation, the main results indicate that innovation has taken importance in identifying the factors of competitiveness, which affects the momentum of activities and generates value in the growth and development of economic sectors of companies that is becoming increasingly competitive [4].

II. Literature Review

Identified the wide variety of aspects that affect the organization of the company, such as planning and production control determining the anticipated way to raise objectives and scenarios that allow to organize and intend to achieve the established programs, to verify the work process or management in an organized and fast way, this control arises from the need to monitor the operations of productive process as the operation of different activities of the quality of products and services obtained [5].

The author in [6] raised a system that allows research to search, check and establish the objectives outlined, innovation generates competitive advantage to seek strategy that will enhance the sector of a not too distant future to develop the quality of resources and have the strategic management that leads organizations to be one step ahead of competitors [7].

The author in [8] manifests a research plan to analyze the state of available activities and resources and innovate performance in the relationship of obtaining competitive ability [9]. The internal control system of innovation processes allows to implement the improvement of business management of the sector [10]. The tool to achieve a better development of this internal control of research and innovation, establishes

activities that ensure safeguarding the identity assets of the information of the result records, to take indispensable action in the continuous improvement of the company, so that it identifies the activities that generate greater added value and be more competitive [11].

The author in [12] indicated, the entrepreneurial capacity of the productive sector has become the main essential element of the sustainable competitiveness of the sector, so the ability to innovate has allowed to achieve sustainable development, which is a major factor of a new environment to know better methods that is to innovate and improve these establishing ways that cause changes in products, processes and organizational systems, for such reasons companies must become aware in innovation that allows to generate competitive advantage [13] (Fig. 1).

A. *Competitiveness and Innovation in the Perspective of Business Development*

The author in [14] indicates the application, is oriented in developing and implementing standard controls with features applicable to the environment that interact the activities or processes in strategic plans.

The author in [15] mentioned that innovation is increasingly relevant in the agenda of entrepreneurs, the organizational model has been articulated from the control of production that allows to seek precisely the strategy for decision making in planning and innovate and be more competitive, in a changing world it has become a commonplace that innovation is key in the competitive construction in companies and sectors to continuous growth in innovation aspect, which allows to generate research and development centers or productive dynamics to realize and determine the main systemic approach [16].

The author in [17] stated to product or service innovation that allows the high degree of improvement with respect to the characteristics of implementing appropriate and technological methods in production, distribution or service with respect to enterprise resource planning that automates business practices in operational or productive aspect of the company, allowing to implement a new method of organizations applying business practice [18].

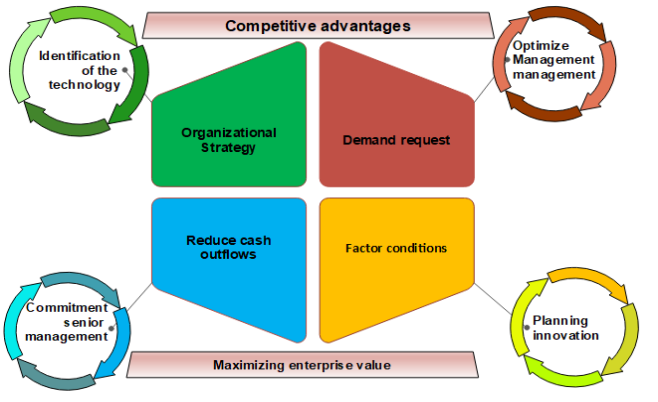


Fig. 1. Key Factors to Generate Competitive Advantage.

The author in [19], pointed out, the emergence of new marketing channels, has enabled the revolution of the application of technology and driven the change of market behavior that allows to obtain the business strategy, in achieving competitive advantage in innovation to learn the right mechanisms and put them into practice, the value of the premise that allows differentiation and positioning against the competition. Definitely [20]. Aspects allow to intervene in business development to have a more accessible notion and obtain results that bring benefits for research and knowledge to facilitate innovation in the generation of new ideas, novel and competitive products [21]. Therefore, it is shown in Fig. 1, competitive advantages, such as technology identification, optimization, and innovation are to obtain favorable results.

Translated with www.DeepL.com/Translator (free version), the author in [22] mentions the innovation of the entrepreneur as an engine of sustainable development, a key factor of the continuous growth model and the application of technology for the introduction of market goods, methods and generation of new raw materials allows to create from an idea, invention or recognition of a need for product development. The successful introduction in the market of new products or management and organizational techniques allows to innovate or create added value, new and significant [23].

The author in [24] focused on innovate, allows to create new value in the market on the premise of competitiveness that depends on the ability to innovate and improve to achieve competitive advantage, through technological innovation in the implementation of ideas, generate value to the product, consumer satisfaction and economic growth [25].

The author in [26] defines an open model of innovating and planning management systems, in which to implement a new commercial market, allows to improve product design in the presentation and development in an appropriate manner in the formulation of the response that allows to route the successful implementation of concrete result, there are factors that condition the economic environment by globalization, which arises the new competitiveness that surrounds us in the company and unstoppable in the growth of technological development [27].

B. *Challenges of Innovation for Sustainable Business Development*

The author in [28] stated, investigating the capabilities that allows companies to drive to be more competitive in innovation to become a strategic level, which we can group into two maneuvers that allow to improve productivity through cost reduction and highly innovative products.

Defined the environment of the company with economic growth of technological development and innovation, which forces companies to be more competitive, in innovation should consider the need to ensure business sustainability to productivity to maintain and improve competitiveness based on innovation, allows to design strategies of plans in differentiation to realize current new product designs and seek the new attributes of requirements and services requested by users [29].

The author in [30] proposed to investigate and invest the mechanism of adaptation of customer need that allows to create and develop a brand image to take of reference of the value chain of the company, the different elements that integrate allow to innovate and reorient the market modeling with the business objective, we can innovate periodically performing new designs of the products by investing in research, infrastructure, equipment and technology to implement the comprehensive development of a new competitive corporate culture.

The author in [31] mentioned that organizational research allows to determine the concentration and establish an innovative culture of different type of scope, in that way to reach and determine tools that help in the capabilities to create and innovate in the organizational method, allowing the resolution of the problem to synthesize and process information about certain instance; the challenge or problems in a visual way allows the exploration of innovative solutions [32].

The author in [33] details in the method of solving questions lightly that we must determine, the process of approaching each participant in assigning the premise on what we raise to create the reality to explore all possible variables, with the achievement of specific objectives that allows to know different scenarios to enrich the structure and identify and solve problems; in which to discover the implicit or explicit needs of the user to present alternatives and solve problems by generating a series of solutions through innovative prototypes.

The author in [34] proposed to provide information opportunities to generate new ideas through innovative thinking, and show the method of generating a business model, which allows to organize the development of an innovative and viable product model consisting of the methodology to assist in improving efficiency and to visualize the effectiveness of sustainable development.

As shown in Fig. 2, the structure of linear thinking allows to apply areas of thinking, observe and identify the innovative idea that comes from the search for opportunities in the environment of the ability to perceive the needs of the organization in concentrating a design to collect optimal service with innovative thinking to understand and focus on providing solutions to problems and provide opportunities to concentrate on the activity of discovering and getting more information [35].

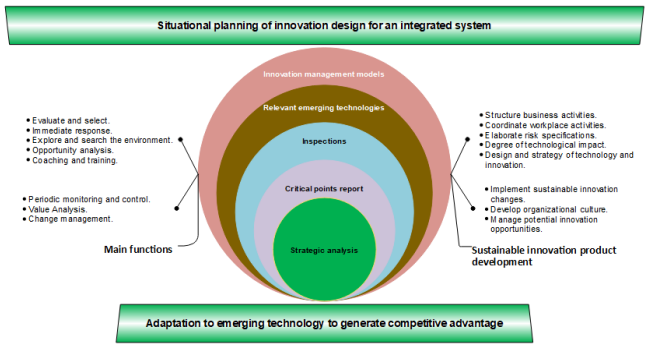


Fig. 2. Parameters of Technological Innovation Management.

The author in [36] detailed to the scheme that allows to demonstrate competitive advantage, constitute the key to improvement and sustainability that provides the perspective based on the importance of product innovation processes, productivity and sustainable development in competitive sectors and segments of companies that have the ability to improve and innovate to create and maintain capacity in a new approach to how to compete, in which it is to detect and discover segment of product and process characteristics to expand and refine the source of competitive [37].

The author in [38] manifests the dynamics of the innovative system of the first instance, promotes sustained investments as the best mechanism to circumvent the differences in productivity generated with a focus on gaining a competitive advantage, which allows to emerge the adversity of change and innovation of variables or key factors to achieve business success with a focus to be able to compete in the market, which lies in the ability of the company to innovate and maintain a competitive advantage [39].

III. Methodology

The present research encompasses a descriptive type design study of methodological approach that collects data from different aspects of the innovation system as a source of market competitiveness for small businesses, allows to perform a measurement analysis of improvements to implement and describe the behavior of the study variables, Therefore, a design is required to expose the thorough form in the studies to manifest the knowledge from the point to implement through structural equations for business development, this study aims to make a development for decision making through research that allows building elements that help identify the characteristics of analysis and diagnosis of all the factors of innovation to generate greater competitiveness [40].

The research presents a quantitative and non-experimental approach to collect data and information on small enterprises in Metropolitan Lima, allowing an analysis and diagnosis, thus concentrating on factors and variables that allow research and innovation to improve competitiveness. The units of analysis are determined in the diagnostic systems in order to propose immediate solutions that help to fulfill the functions for which they have been acquired, by prioritizing them [41].

During the research process, the redesign of the competitiveness process should be considered for continuous improvement through sustainable business development, allowing to determine and establish control parameters for this improvement, knowing any event that affects the performance of the markets to reduce the risks involved in taking measures to minimize possible losses and incorporate best practices for the implementation of methods, methodology, procedures for continuous improvement to improve the business economic development that will enable the most effective and efficient use in the available production that will allow obtaining the greatest possible amount of goods and services at a lower cost [42]. Establishing research and innovation actions to determine and minimize risks in operations, allows to establish the management to expand, optimize productivity and sustainable development of the company, in the research proposed in this document optimal decisions are made to expand capabilities in

business development that evaluates the management in a comprehensive, systematic way, taking advantage of the opportunities presented by the company [43].

The degree of competitiveness of the company is a key factor in the development for making decisions in an efficient and timely manner that allows to evaluate different areas of the company, thus being able to develop the proposed management model and the priority importance of evaluating the situation of the company to detect existing problems, establish priorities that allow to know the strengths and weaknesses and measure the productive performance of its economic activity.

IV. Results

The data analysis according to the structural methodology of the elaboration of structural equations proposed in the research methodology; starts with the reliability analysis of the instrument, evaluation of the factorial analysis, exploratory factorial analysis, confirmatory factorial analysis and elaboration of the structural equations model that performs the descriptive research of factors that obtains in carrying out the hypothesis testing ending with the results obtained. To obtain the Cronbach's Alpha with a coefficient 0.924, allows measuring the reliability of a measurement scale for the 39 items and a Cronbach's Alpha based on standardized items with a value of 0.924, by which, it indicates that there is an excellent level of internal consistency of instrument scale that exceeds the value of 0.9, explained in methodology of the present research can be observed in Table I.

TABLE I. Reliability Analysis and Diagnosis of the Instrument

Reliability statistics		
Alfa de Cronbach	Cronbach's alpha based on standardized items	N of elements
.924	.924	39

A. Item Reliability Analysis of Independent Variable

For the case of the independent variable innovation management model, the Cronbach's Alpha coefficient was obtained based on the standardized elements with a value of 0.918 higher than 0.9, obtaining an excellent coefficient in the indicators, as shown in Table II.

TABLE II. Analysis and Diagnosis of the Reliability of the Indicators of the Innovation Management Model

Reliability statistics		
Alfa de Cronbach	Alfa de Cronbach based on standardized items	N elements
.918	.918	20

B. Dependent Variable Item Reliability Analysis

It will allow to determine the degree of relationship between them, it is advisable to perform an individual analysis, for each variable and the dimension detailing the results obtained, by means of SPSS software version 25. For the case of the dependent variable source of business competitiveness of the Industrial SMEs, the Cronbach's Alpha coefficient based on the standardized elements 0.885 higher than 0.8 has been obtained, obtaining a good level coefficient in the indicators, which can be observed in Table III.

TABLE III. Analysis and Reliability Diagnosis of the Indicators of the Source of Business Competitiveness of Industrial SMEs

Reliability statistics		
Alfa de Cronbach	Alfa de Cronbach based on standardized items	N elements
.885	.885	19

C. Bartlett's Test of Sphericity Contrast Test

In the overall sufficiency analysis of the innovation management model as a source of business competitiveness of Industrial SMEs of the data evaluated with the data taken, by SPSS software version 25. With the KMO index of 0.906 given is considered an excellent value as seen in Table IV.

TABLE IV. KMO and Bartlett's Test for Integral Results

KMO and Bartlett's test		
Kaiser-Meyer-Olkin measure of sampling adequacy		
Bartlett sphericity test	Chi-square approximation	14186.893
	gl.	741
	Sig.	.000

D. Total Variance Explained

Considering the eight factors obtained in the explained variance table that should consider the existence of factors that are evidently that could not contribute the large extent of the structural equation model, for which, the use of exploratory result is required to be able to start the confirmatory analysis that could contribute to the theory of the research can be observed in Table V.

As can be seen in Fig. 3, the standardized confirmatory factor analysis values can be considered as the number of standard deviations by which the adjusted residuals differ from the zero-valued residuals, which will be associated with a perfect fit model.

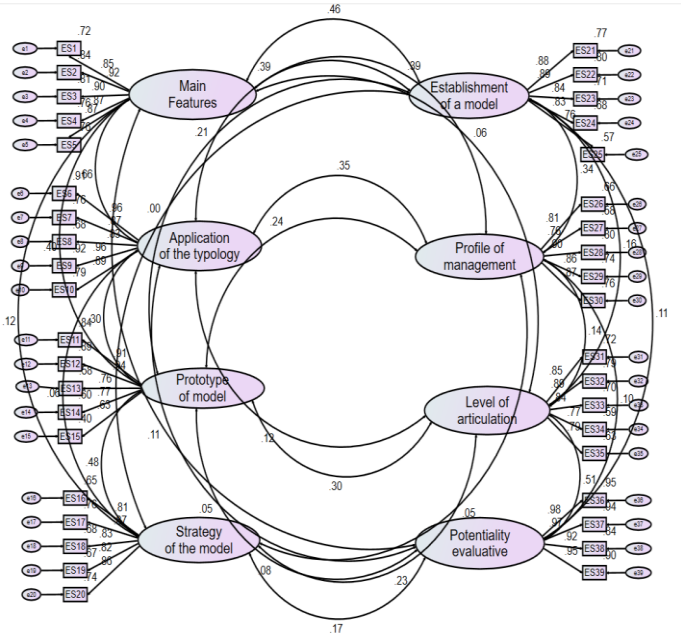


Fig. 3. Standardized Confirmatory Factor Analysis.

TABLE V. INTEGRAL EXPLAINED VARIANCE

Total, variance explained									
	Initial eigenvalues			Sums of loads squared by extraction			Sums of loads squared by rotation		
Component	Total	% of variance	% accumulated	Total	% of variance	% Accumulated	Total	% of variance	% accumulated
1	10.558	27.073	27.073	10.558	27.073	27.073	4.305	11.039	11.039
2	5.857	15.017	42.090	5.857	15.017	42.090	4.018	10.304	21.343
3	4.339	11.124	53.214	4.339	11.124	53.214	3.974	10.189	31.532
4	2.833	7.264	60.478	2.833	7.264	60.478	3.907	10.017	41.549
5	2.527	6.480	66.958	2.527	6.480	66.958	3.881	9.953	51.502
6	2.143	5.495	72.453	2.143	5.495	72.453	3.847	9.863	61.365
7	1.703	4.366	76.819	1.703	4.366	76.819	3.684	9.447	70.812
8	1.292	3.312	80.131	1.292	3.312	80.131	3.634	9.319	80.131
9	.654	1.677	81.808						
10	.512	1.312	83.120						
Extraction method: principal component analysis									

The results of the measurement scale adjustments of the confirmatory factor analysis model, the variables of the present research, it can be verified that most of them comply with a good adjustment of the model, specifically CFI=0.950, IFI= 0.950, TLI=0.945, NFI = 0.906 all complying with the norms with values higher than 0.90 and RMSEA= 0.053 lower than 0.08 complying with the specification of X^2 can be seen in Table VI.

The results of the fit measures of the confirmatory factor analysis model are detailed for the construction of the model and the results of the loadings of the unstandardized estimators of the structural equation model are also shown.

Fig. 4 shows the diagram of the values of the unstandardized loadings, it can be seen that it has a value greater than 0.5 from the latent variable to the observed variable; therefore, it presents an acceptable factorial loading.

TABLE VI. INDICATORS OF MODEL FIT MEASUREMENT

Statistics of model fit measurement indicators		
Absolute Adjustment Measures		
X^2	Chi-square and significance level (p)	1379.236
GFI	Goodness of Fit index	0.838
RMSEA	Root mean square error of approximation	0.053
NCP	Nocenttrality Parameter	705.236
RFI	Relative fit index	0.897
ECVI	Expected cross-validation index	4.278
RMR	Root mean square residual	0.043
Incremetal Adjustment Measures		
AGFI	GFI adjusted goodness of fit index	0.812
CFI	Comparative fit index	0.950
IFI	Incremental fit index	0.950
TLI	Tucker Lewis index	0.945
NFI	Normed fit index	0.906
Parsimony adjustment measures		
X^2	Normalizada X^2 / d.f.	2.046
PNFI	Parsimony normed fit index	0.824
PGFI	Parsimony goodnee of fit index	0.724
Others		
AGFI	Corrected goodness-of-fit index	0.812

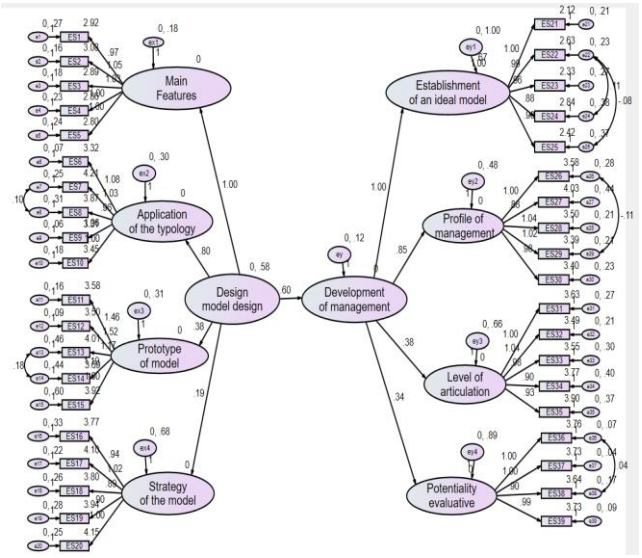


Fig. 4. Diagram of Non-Standardized Load Values.

The model with the correlations created in the described errors, by suggestion shows 693 degrees of freedom, which is obtained from a different number of parameter equal to 780 and a different number to parameter to estimate 87 obtaining 693 of difference (780-87=693), the changes that were made in this model in a definitive way, the structure equation that shows the values of adjustment measures, the results of the adjustment measures of the model of the structural equations of the variables of the present investigation are obtained, it will be possible to verify that it fulfills the majority in good adjustment of model that specifies CFI= 0. 951, IFI= 0.951, TLI=0.947, NFI=0.907, all complying with the norm of being equal to or higher than 0.9 and including X^2 = 2.003, which has a value between 2 and 5, by which, the model allows to obtain a good quality can be appreciated in Table VII.

Fig. 5 shows the diagram of improved specific values with unstandardized loadings. It can be seen that it has a value greater than 0.5, from the latent variable to the observed variable, therefore, it presents an acceptable factorial loading.

TABLE VIII. STATISTICS OF MODEL FIT MEASUREMENT INDICATORS

Statistics of model fit measurement indicators		
Absolute adjustment measures		
X^2	Chi cuadrado y nivel de significancia (p)	1376.142
GFI	Goodness of Fit index	0.840
RMSEA	Root mean square error of Appoximation	0.052
NCP	Nocenttrality Parameter	689.142
RFI	Relative fit index	0.899
ECVI	Expected cross-validation index	4.199
RMR	Root mean square residual	0.096
Incremental adjustment measures		
AGFI	GFI Adjusted goodness of fit index	0.818
CFI	Comparative fit index	0.951
IFI	Incremental fit index	0.951
TLI	Tucker Lewis index	0.947
NFI	Normed fit index	0.907
Parsimony adjustment measures		
X^2	Normalizada X^2 / d.f.	2.003
PNFI	Parsimony normed fit index	0.841
PGFI	Parsimony goodnee of fit index	0.739
Others		
AGFI	Índice de bondad de ajuste corregido	0.818

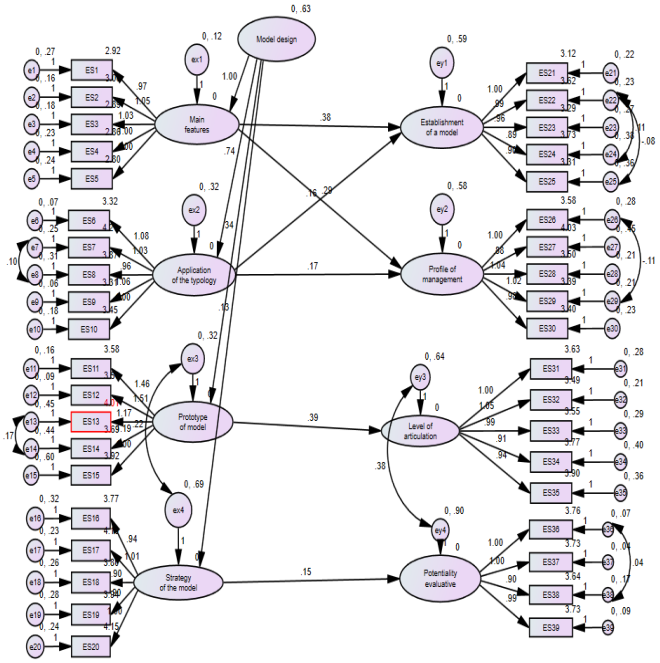


Fig. 5. Data Taken by AMOS software Version 24.

The detail of structural equations of the specific model considering X1: Main characteristics of the business management of industrial SMEs; X2: Generate the typology of industrial SMEs; X3: Prototype of integral system model; X4: Strategy of the proposed integral model through its business application; Y1: Establishment of an ideal model; Y2: Profile of business management; Y3: Level of articulation of variables and cause-effect relationships; Y4: Evaluative potentiality of its parameters in an exact manner, can be seen in Table VIII.

TABLE IX. SPECIFIC EQUATION MODEL

Integral system model plan indicators
X1 = 0.917 (X) + 0.119
X2 = 0.721 (X) + 0.323
X3 = 0.429 (X) + 0.321
X4 = 0.120 (X) + 0.688
Y1 = 0.373 (X1) + 0.589
Y2 = 0.301 (X1) + 0.582
Y3 = 0.292 (X3) + 0.640
Y4 = 0.129 (X4) + 0.900

E. Specific Hypothesis

In the specific hypothesis test of the result base of the specific theoretical model of structured equations, in the research methodology allows to obtain the data for a better interpretation of the specific hypothesis test, it will be based on the equation of the research methodology of the equations of the following table, it can be observed in Table IX.

TABLE X. REGRESSION FOR HYPOTHESIS TESTING OF SPECIFIC STRUCTURAL EQUATION

Hip.	Y = λX + Error		**λ**	**C.R.**	**P**
	(λ Non-standardized)		**Standardized**		
H1	Y1 ← 0.353 (X1) + 0.340		0.051	6.677	***
H2	Y2 ← 0.173 (X2) + 0.165		0.050	3.302	***
H3	Y3 ← 0.246 (X3) + 0.331		0.075	4.413	***
H4	Y4 ← 0.180 (X4) + 0.210		0.062	3.37	***

From the results obtained from the regression line of the structural equation model H1, H2, H3 and H4, a positive standardized loading greater than 0 is observed, validating the base of the equation $p<\alpha$ (α=0.05) can be verified, for which, the hypothesis (H0) is rejected and the alternative hypothesis (Ha) is accepted. It allows specifying if there is influence between the independent and dependent variables.

V. CONCLUSION

The main contribution of the research allows building the design of the innovation management model as a source of business competitiveness of industrial SMEs, which is strictly related to the hypothesis in alignment with the objectives of the problems posed in this research.

a) H1 was tested, therefore, it can be stated that the main characteristics of the business management of industrial SMEs and the establishment of an ideal model there is a highly significant direct relationship between the variables.

b) H2 was tested, therefore, it can be stated that the application of the typology of industrial SMEs improves the

profile of business management. There is a highly significant direct relationship between the variables.

c) H3 was tested, therefore, it can be affirmed that the prototype of the integral system model improves the level of articulation of the variables of cause-effect relationships. There is a very significant direct relationship between the variables.

d) H4 was proved, therefore, it can be affirmed that the strategy of the proposed integral model through its business application improves in evaluative potentiality of its parameters in an exact way. There is a very significant direct relationship between the variables.

REFERENCES

[1] Amis, J., Mair, J. y Munir, K. (2020). The Organizational Reproduction of Inequality. Journal Academy of Management Annals, 14(1), 195-230. https://doi.org/10.5465/annals.2017.0033.

[2] Ascani, A. y Iammarino, S. (2018). Multinational enterprises, service outsourcing and regional structural change. Cambridge Journal of Economics, 42(6), 1585-1611. https://doi.org/10.1093/cje/bey036.

[3] Bailey, D., Corradini, C. y Propis, L. (2018). 'Home-sourcing' and closer value chains in mature economies: the case of Spanish Manufacturing. Cambridge Journal of Economics, 42(6), 1567-1584. https://doi.org/10.1093/cje/bey020.

[4] Barón, E., Garcia, C. y Sanchez, S. (2021). La inteligencia de negocios y la analítica de datos en los procesos empresariales. Revista Científica de Sistemas e Informática, 1(2), 37 - 53. DOI: https.doi.org.10.51252.rcsi.v1i2.167.

[5] Basán, N. (2019). Modelos avanzados de optimización para la gestión eficiente de procesos de producción. [Tesis doctoral, Universidad Tecnológica Nacional Santa Fe]. Repositorio: Universidad Tecnológica Nacional Santa Fe.

[6] Beckert, J. (2019). Markets from meaning: quality uncertainty and the intersubjective construction of value. Cambridge Journal of Economics, 44(2), 285-301. https://doi.org/10.1093/cje/bez035.

[7] Belabed, C., Theobald, T. y Treeck, T.(2017). Income distribution and current account imbalances. Cambridge Journal of Economics, 42(1), 47-94. https://doi.org/10.1093/cje/bew052.

[8] Belfrage, C. y Kallifatides, M. (2018). Financialisation and the New Swedish Model. Cambridge Journal of Economics, 42(4), 875-900. https://doi.org/10.1093/cje/bex089.

[9] Benson, J. (2017). Environmental law & the limits of markets. Cambridge Journal of Economics, 42(1), 215-230. https://doi.org/10.1093/cje/bex027.

[10] Berman, S. (2012). Digital transformation: opportunities to create new business models. Strategy & Leadership, 40(2), 16-24. https://doi.org/10.1108/10878571211209314.

[11] Braunstein, E., Bouhia, R. y Seguino, S. (2019). Social reproduction, gender equality and economic growth. Cambridge Journal of Economics, 44(1), 129-156. https://doi.org/10.1093/cje/bez032.

[12] Bustamante, C., Bustamante, M. y Morales, D. (2017). Inteligencia de negocios y su incidencia en las organizaciones. INNOVA Research Journal, 2(8),159-173. DOI: https_doi.org_10.33890_innova.v2.n8.1.2017.360.

[13] Caffarel, Y. (2018). The nature of heterodox economics revisited. Cambridge Journal of Economics, 43(1), 527-540. https://doi:10.1093/cje/bey054.

[14] Cameron, K. y Quinn, R. (2011). Diagnosing and changing organizational culture. San Francisco, EEUU: Jossey Bass.

[15] Carabelli, Anna y Cedrini, M. (2018). Great Expectations and Final Disillusionment: Keynes, ''My Early Beliefs'' and the Ultimate Values of Capitalism. Cambridge Journal of Economics, 42(5), 1183-1204. https://doi.org/10.1093/cje/bey017.

[16] Casas R. (2019). Optimización del control de los procesos de operación y mantenimiento para una empresa de telecomunicaciones. [Tesis de grado, Universidad Nacional del Callao]. Repositorio: http://repositorio.unac.edu.pe/handle/20.500.12952/3429.

[17] Cohen, D., Asín, E., Lankenau D. y Alanis, D. (2004). Sistemas de información para los negocios: Un enfoque para la toma de decisiones (3 ed). México, México: McGraw-Hill/Interamericana.

[18] Coles, J., Li, Z. y Wang, A. (2017). Industry Tournament Incentives. Review of Financial Studies, 31(4), 1418-1459. https://doi.org/10.1093/rfs/hhx064.

[19] Cordero, D. y Rodríguez, G. (2017). La inteligencia de negocios: una estrategia para la gestión de las empresas productivas. Revista Ciencia UNEMI, 10(23), 40-48. DOI: 2528-7737.

[20] Corvalán, J. (2018).Inteligencia artificial retos, desafíos y oportunidades Prometea: la primera inteligencia artificial de Latinoamérica al servicio de la justicia. Revista de Investigações Constitucionais, 5(1), 295-316. DOI: 10.5380/rinc.v5i1.55334.

[21] Dafermos, Y. (2017). Debt cycles, instability and fiscal rules: a Godley–Minsky synthesis. Cambridge Journal of Economics, 42(5), 1277-1313. https://doi.org/10.1093/cje/bex046.

[22] Delfín, L. y Acosta, M. (2016). Importancia y análisis de desarrollo empresarial. Pensamiento y Gestión, 40(1), 184-202. https://doi.org/10.14482/pege.40.8810.

[23] Fernández, M. (2017). Challenges of economic globalization. Revista de relaciones Internacionales, Estrategia y Seguridad, 12(1), 23-50. http://dx.doi.org/10.18359/ries.2462.

[24] Flore, H., Pérez, G., Gioia, L. y Medina, J. (2018). Metodología de optimización de procesos industriales relacionando las inversiones con los costos operativos. Revista Portal de la Ciencia, 1(14), 87-95. DOI: https://doi.org/10.5377/pc.v0i14.6641.

[25] Hassan, T. y Tahoun, A. (2017). The Power of the Street: Evidence from Egypt's Arab Spring. Review of Financial Studies, 31(1), 1-42. https://doi.org/10.1093/rfs/hhx086.

[26] Heredero, C., Agius, J., Romero, S. y Salgado, S. (2012). Organización y transformación de los sistemas de información en la empresa. Madrid, España: ESIC Editorial.

[27] Jacobs, F. (2007). Enterprise resource planning (ERP) A brief history. Journal of Operations Management, 25(2), 357-363. https://doi.org/DOI:10.1016/j.jom.2006.11.005.

[28] James, G., Chad, A. y Hannes L.(2019). Taking stock of moral approaches to leadership: an integrative review of ethical, authentic, and servant Leadership. Journal Academy of Management Annals, 13(1), 148-187. https://doi.org/10.5465/annals.2016.0121.

[29] Kuys, B. Koch, C., Renda, G. (2021).The Priority Given to Sustainability by Industrial Designers within an Industry 4.0 Paradigm. Sustainability, 14(1), 1-17. https:doi.org_10.3390_su14010076.

[30] Mielgo, N., Montes, J. y Vázquez, C. (2007). Cómo gestionar la innovación en las Pyme. Madrid, España: Netbiblo.

[31] Mintzberg, H. (2004). Una visión crítica de la dirección de empresas y la formación empresarial. Barcelona, España: Ediciones Deusto.

[32] Morris, J., Wang, W., Shah, D. Plaisted, T. y Hansen, C. y Amirkhis (2022). Expanding the design space and optimizing stop bands for mechanical metamaterials. Materials and Design, 216(2022),1-13. DOI: https_doi.org_10.1016_j.matdes.2022.110510.

[33] Muhammad, Y., Mohd. K., y Siti M.(2022) Modified teaching learning based optimization for selective harmonic elimination in multilevel inverters. Ain Shams Engineering Journal 13(2022),3-6. DOI: https://doi.org/10.1016/j.asej.2022.101714.

[34] Muñoz, H., Osorio, M. y Zuñiga, P. (2016). Inteligencia de los negocios: Clave del Éxito en la era de la información. Revista Clío América, 10(20), 194-211. DOI: http://dx.doi.org/10.21676/23897848.1877.

[35] Ottati, G. (2017). Marshallian Industrial Districts in Italy: the end of a model or adaptation to the global economy? Cambridge Journal of Economics, 42(2), 259-284. https://doi.org/10.1093/cje/bex066.

[36] Schwertner, K. (2017). Digital transformation of business. Trakia Journal of Sciences, 15(1), 388-393. https://doi:10.15547/tjs.2017.s.01.065.

[37] Senoner, J., Netland, T., y Feuerriegel, S. (2021). Using Explainable Artificial Intelligence to Improve Process Quality: Evidence from Semiconductor Manufacturing. Magazine Management Science. 0(0), 1-20 DOI: https//doi.org/10.1287/mnsc.2021.4190.

[38] Venegas, R.(2021) Aplicaciones de Inteligencia Artificial Para La Clasificación Automatizada de Propósitos Comunicativos En Informes de Ingeniería. Revista Signos, 54(107), 942–970. DOI: 10.4067/S0718-09342021000300942.

[39] Viteri, C. y Murillo, D. (2021). Inteligencia de Negocios para las Organizaciones. Revista Arbitrada Interdisciplinaria Koinonía, 6(12), 304-333. DOI: http://dx.doi.org/10.35381/r.k.v6i12.1291.

[40] Walas, F. y Redchuck, A. (2021) IIoT/IoT and Artificial Intelligence/Machine Learning as a Process Optimization Driver under Industry 4.0 Model. Journal of Computer Science & Technology, 21 (2),170-176, 2021. http://dx.doi.org/10.24215/16666038.21.e15.

[41] Ziyadin, S., Suieubyeva S. y Utegenova, A. (2020). Digital Transformation in business. Revista ISCDTE, 84, 408-415. https://doi.org/10.1007/978-3-030-27015-5_49.

[42] Ramos, N., Fernández, A. y Almodóvar, M. (2020). El impacto de las TIC en el rendimiento de la Pyme: estado actual de la cuestión. Revista espacios, 41(25), 387-403. ISSN: 0798-1015.

[43] Cutipa, A.; Escobar, F., Anchapuri, M., Valreymond, D. (2020). La intensidad de innovación y la competitividad de micro y pequeñas empresas exportadores de artesanía textil. Revista Escuela de Administración de Negocios, 89(1), 155-176. DOI: https://doi.org/10.21158/01208160.n89.2020.2848.

A Cross Platform Contact Tracing Mobile Application for COVID-19 Infections using Deep Learning

Josephat Kalezhi[1]
Department of Computer Engineering
The Copperbelt University
Kitwe, Zambia

Christopher Chembe[3]
Department of Computer Science
ZCAS University
Lusaka, Zambia

Francis Lungo[5]
School of Social Sciences
Mulungushi University
Kabwe, Zambia

Mathews Chibuluma[2]
Department of Information Technology/Systems
The Copperbelt University
Kitwe, Zambia

Victoria Chama[4]
Department of Computer Science and Information Technology
Mulungushi University
Kabwe, Zambia

Douglas Kunda[6]
Department of Computer Science
ZCAS University
Lusaka, Zambia

***Abstract*—The COVID-19 pandemic has remained a global health crisis following the declaration by the World Health Organization. As a result, a number of mechanisms to contain the pandemic have been devised. Popular among these are contact tracing to identify contacts and carry out tests on them in order to minimize the spread of the coronavirus. However, manual contact tracing is tedious and time consuming. Therefore, contact tracing based on mobile applications have been proposed in literature. In this paper, a cross platform contact tracing mobile application that uses deep neural networks to determine contacts in proximity is presented. The application uses Bluetooth Low Energy technologies to detect closeness to a Covid-19 positive case. The deep learning model has been evaluated against analytic models and machine learning models. The proposed deep learning model performed better than analytic and traditional machine learning models during testing.**

***Keywords*—Contact tracing mobile application; coronavirus; COVID-19; deep neural networks**

I. Introduction

In March 2020, the coronavirus disease (COVID-19) was declared a pandemic by the World Health Organization [1]. Since then, relentless efforts were put in place by several nations to understand the virus and how to contain the pandemic. One of the promising approaches is digital contact tracing [2]. Contact tracing has been used to follow the pattern of networks for an individual or population infected by an infectious disease. In the past, contact tracing has been used to combat sexually transmitted diseases, severe acute respiratory syndrome (SARS) and other invading pathogens [3]. Traditionally, there have been various contact tracing models including Individual-based simulation models; Pair approximation models; Models based on branching processes; and Phenomenological approaches [4]. These come with various challenges such as the inability for stochastic simulation-based models to be analysed analytically. Other challenges are associated with contact structure itself, backward- and forward tracing, identification of Super-spreaders, endemic equilibrium and efforts required for contact tracing [4]. The effectiveness of various contact tracing mechanisms has been presented by Klinkenberg et al [5].

In recent years, digital contract tracing has been championed to supplement the deficiencies introduced by traditional contact tracing. For example, a mobile contact tracing application was developed to trace and monitor Ebola epidemic in Northern Sierra Leone and proved to be effective [6]. Similarly, Sacks et al. developed a smartphone based mHealth application using CommCare and business intelligence software Tableau to assist in contact tracing of Ebola epidemic in Guinea [7]. Swanson et al. [8] gives details on the performance of contact tracing in Liberia during the 2014 to 2015 epidemic. Other uses of mobile phones in contact tracing have been used in tracing the spread of Tuberculosis (TB) [9]. Furthermore, the outbreak of COVID-19 and subsequent declaration as pandemic has seen a proliferation of mobile applications aimed at contact tracing to help combat COVID-19 [10] [11]. These applications have proved effective in tracing contacts in order to contain the pandemic [2].

In early days of COVID-19 pandemic, Singapore developed a COVID-19 contact tracing mobile app called "TraceTogether" [12]. This app uses Bluetooth technology to facilitate contact tracing. The app further uses the received signal strength indicator (RSSI) values detected from other mobile devices for distance estimation. The detected RSSI values are then compared against calibrated RSSI values to determine distance between mobile devices. The app notifies users when they are exposed to COVID-19 and are in close contact to other users who are using the same app. It also allows users to access their COVID-19 health status. The app uses a BlueTrace protocol to preserve privacy. The reference

implementation of the BlueTrace is referred to as OpenTrace [13]. Through the Ministry of Health, the app provides guidelines on how to avoid getting infected. In case one tests positive for COVID-19, the data is then shared with Ministry of Health. The Bluetooth data is kept on the phone no longer than 25 days. Despite being useful, the application has been criticised for the potential in being exploited in undertaking criminal investigations by the police.

Similarly, a National Health Service COVID-19 app was developed in England and Wales [14]. The app gave an option to users to enable contact tracing. It was reported that the effectiveness of the app towards reducing infections was dependent on the number of users. Bluetooth RSSI values were used to estimate distance between two close devices. Several RSSI values were taken and then used to determine the distance. Another contact tracing app called Immuni was developed in Italy by the Ministries of Health and Technological Innovation in 2020 [15]. The app used Bluetooth technology to facilitate contact tracing. The distinguishing functionality of the app was the absence of a centralized database to manage contact tracing. For users willing to use the app, they would download the data to support contact tracing in their smartphones and a decision made locally in case the users were exposed. Privacy concerns were also addressed in the development of the app in line with the national laws.

A contact tracing app called "Radar Covid" was developed in Spain in 2020 [16]. Users of the app received notifications when they were in close contact with a COVID-19 positive person. The app used Bluetooth low energy technology to detect if the user is in close proximity to a positive case. The app was voluntary and addressed the privacy and security concerns of users. When the user test positive, the user is presented with an anonymous code that can be entered into the app voluntarily. This in turn facilitated the notification of other users in case they had been in contact with a positive case. Despite being useful the app had a vulnerability that allowed attackers to use fake identities.

In 2020 Apple and Google joined forces to develop application programming interfaces (APIs) to facilitate contact tracing [17]. The technology adopted was Bluetooth owing to its availability in virtually every mobile device. The APIs are used to detect contacts typically within two meters for a period exceeding 15 minutes [18]. As pointed out in [17], the APIs have addressed privacy concerns by preventing access to user profiles by health authorities.

Other technology-based contact tracing mechanisms to fight COVID-19 has been employed previously. For example, a combination of machine learning classification algorithm and data obtained from Wi-Fi signals from users was proposed to determine when two users sharing the same physical space can inform exposure [19]. In [20], a framework based on IoT was proposed for contact tracing. They incorporated symptom-based detection ignored in other tracing models to confirm COVID-19 cases. The work proposed by Sahraoui et al used online social network to trace COVID-19 infections [21].

Despite the many works presented on contact tracing, there is more to be done for digital contact tracing to be appreciated in future. One challenging issue is privacy and protection of user data. In this paper, we present a cross platform contact tracing application that uses Bluetooth Low Energy Generic Attribute Profile (GATT) framework and deep neural network to determine and inform users of exposure to COVID-19. GATT framework is used to mitigate the many concerns regarding data privacy and protection. The data exchanged between Bluetooth devices embedded with GATT framework is encapsulated thereby authorizing only intended recipient. The deep neural network is applied to predict the distance between communicating devices. Using GATT framework to encapsulate packets exchanged between devices and deep neural network to predict distance has not been presented in literature. Hence the proposed approach presents novelty and contribution of this paper.

The rest of the paper is structured as follows. In Section II we review the literature related to this work. Section III presents the Received Signal Strength Indicator (RSSI) obtained from Bluetooth Low Energy (BLE) devices. In Section IV, a logarithmic distance path loss model applied in this work is presented. A decision tree is presented in Section V. The proposed deep neural network model that uses the RSSI is reported in Section VI. A comparison of performances of models is presented in Section VII. Section VIII reports the development of a cross-platform contact-tracing mobile application. The conclusion appears in Section IX.

II. Related Work

In order to aid with the digital contact-tracing process, a number of mobile applications have been developed worldwide [11]. These applications have proved effective in tracing contacts in order to contain the pandemic [2]. However, despite being useful a number of challenges still remain [22]. These include security and privacy concerns by users sharing the data, transparency, the effectiveness of the tracing application, social and cultural issues, legal and ethical issues and many more [23]. Megnin-Viggars et al. [24] identifies other barriers and factors to engaging in contact tracing during an infectious pandemic such as COVID-19. To mitigate some of the challenges and barriers to digital contact tracing, researchers have proposed various solutions. For instance, blockchain technology has been proposed to preserve privacy during contact tracing for COVID-19 pandemic to gain trust by users [25][26][27]. According to [26], it was reported that blockchain technology was able to detect unknown cases of COVID-19. The application was also capable of enabling individuals to use the mobile application to predict the probabilities of being infected. The study paved way for the use of blockchain technology to contain the spread of the epidemic as well as early detection of unknown infections.

The works in [28], proposed a smart contact tracing mobile application that uses Bluetooth Low Energy (BLE) and machine learning techniques. The application determined whether the user was at risk or not depending on whom they came into contact with. An analytic proximity estimation model based on RSSI was particularly used to determine the distance between two devices. Five machine learning classifiers were considered in the estimation of distance

between devices. These were Support Vector Machine, Decision Tree, Naïve Bayes, Linear Discriminant Analysis and K-Nearest Neighbors. It was reported that the Decision Tree classifier yielded the best accuracy compared to other classifiers.

In [29], authors presented a contact tracing application for wearable devices that employs machine learning. The application used BLE for distance estimation whereas machine learning was applied to categorize the risk of possible exposure. Additionally, an appropriate signature protocol was used to guarantee infected user anonymity. The authors studied four supervised-learning classifiers namely Decision Tree, Linear Discriminant Analysis, Naïve Bayes and K-Nearest Neighbours. It was reported that the classifiers performed well and yielded good precision and recall values. The Decision Tree classifier was reported to yield the best performance in terms of precision, recall, accuracy among others.

In [30] authors proposed a BLE application that monitors location patterns of old people indoors. The system relied on RSSI to estimate the positions of these elderly people. To achieve this, BLE beacons were either attached to a person's clothes or worn on wrists. Further the users could also place the beacons in their pockets. The beacons were periodically sending broadcasts. These broadcasts were detected by a number of BLE enabled Raspberry Pi devices that were stationed at fixed known locations. The broadcasts carried the RSSI among others. Each Raspberry Pi then relayed the received data to a server for additional processing. The server ran a machine-learning classifier to determine the location of the person. A path-loss model was used for estimation of distance from RSSI. Further a number of classifiers were used and these are Naïve Bayes, Random Forest, BayesNet, Sequential Minimal Optimization and J48. In overall the performances of classifiers were good for indoor localization.

A dependence of RSSIs on distance for iOS and Android mobile devices was reported in [31]. According to [31], the iOS device was used in that study, the RSSI reached an asymptotic value (where the RSSI appears not to change) earlier than the Android device. It was further reported that the RSSI detected on the Android phone used decreased gradually compared to the iOS device. In terms of temporal RSSI variations, the Android device exhibited more variations compared to iOS. Therefore, according to the study [31], the dependence of RSSI on distance varies between iOS and Android devices.

In a related work [32], an evaluation of a contact tracing mobile application in Norway based on the Google Apple Exposure Notification (GAEN) system was reported. The authors observed variations in Bluetooth attenuation levels, when alerts are generated among others between iOS and Android mobile devices. The Android device was reported to exhibit high variabilities compared to iOS devices. In another related study [33], a number of data mining models have been applied to reveal hidden patterns in patients' data in Zambia. The models include J48 decision classifier, Naïve Bayes, Multilayer Perceptron among others. These models were shown to exhibit good performance compared to baseline results. The COVID-19 cases in Zambia are reported by the Ministry of Health through the Zambia National Public Institute [34]. The contact tracing process used in Zambia prior to this work was manual and time consuming.

It is clear from the works reported above that contact-tracing apps have found their application in mitigating the spread of COVID-19 pandemic. In terms of distance estimation, a number of models have been reported in the literature. These include simplistic path loss models as well as several machine learning models. Nevertheless, no work has considered GATT framework to encapsulate user data in order to mitigate the concerns regarding data privacy and protection. Furthermore, no work has used deep neural network to predict the distance between communicating devices. Thus, in this paper we propose deep learning methods to estimate distance between the devices and the GATT framework to encapsulate the data between communicating devices in order to secure user data. The proposed deep learning model is compared against some models reported in the literature. The deep learning model is further converted to models suitable for use in a contact tracing mobile application. We further developed a cross-platform contact-tracing app for use in Zambia that incorporates these models.

III. Received Signal Strength Indicator

The Received Signal Strength Indicator (RSSI) is a measure of the signal strength detected by a receiving device [35]. The RSSI is manufacturer dependent and can vary even at a fixed separation between the sending and receiving Bluetooth Low Energy (BLE) devices [35]. Factors such as multipath propagation, scattering, shadowing, refraction among others affect BLE signals and this has implications for applications that rely on BLE such as those for contact tracing [36]. The RSSI can be reasonably mapped to a distance from the sending device on iOS devices. However, the quality of RSSI on Android devices varies significantly due to presence of several chip manufacturers [35]. Nevertheless, a variation of this quantity with distance from a sending device can be used as a distance measure [35].

In this work, we relied on BLE RSSI to determine distance between two devices. Measurements were initially taken at several separations between two devices. It was observed that even when the distance was fixed, the RSSI reading was changing as reported in [35]. Measurements were taken at several distances between two devices beginning with an initial separation of 4 m to a final separation of 0.1 m in steps of 0.1m. At each distance, several RSSI measurements were repeated (twenty in this case) and the mean determined. Fig. 1 shows the dependence of the mean RSSI in decibels on actual distance separating two devices in meters. The signal originated from an iOS device (iPhone 7 Plus) and the RSSI was detected on an Android device (Infinix Smart HD). As can be seen from Fig. 1, the mean RSSI follows a particular trend despite the fluctuations as the distance between devices changes.

Similar measurements were undertaken for signals originating from an iOS device (iPhone 11) and detected on another iOS device (iPhone 7 Plus). Measurements were also undertaken for signals originating from an Android device and detected on an iOS device (iPhone 7 Plus). It was observed that

RSSI values decrease in general as the distance increased in all cases. However, variations of RSSI on distance were different for each case. The calibration results of mean RSSI at 2m separation for various devices has been reported for TraceTogether contact tracing app as used in OpenTrace [13]. The results shown in Fig. 1 at a separation distance of 2 m are in agreement with OpenTrace calibration results.

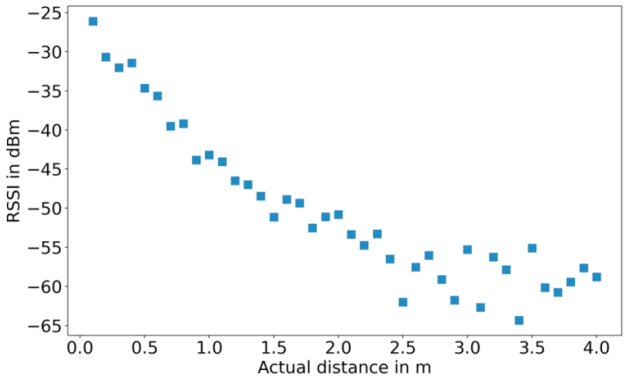


Fig. 1. Dependence of Mean RSSI in dBm on Actual Distance between Two Devices in Metres. The Signal Originated from an iOS Device (iPhone 7 Plus) and Detected on Android Device (Infinix Smart HD).

IV. Logarithmic Distance Path Loss Model

One of the simplest models relating RSSI to distance is the logarithmic distance path loss model [37]. This model is expressed as:

$$RSSI = - 10n \log(d/d_0) + A + X_\sigma \quad (1)$$

where n is an environment dependent path loss parameter, d is the distance from the sending device to the receiving device, d_0 is a distance where the RSSI takes the value A. X_σ is a random variable that follows a Gaussian-distribution. X_σ has zero mean and a variance of σ^2.

Taking the mean of equation (1), one obtains

$$RSSI_{mean} = -10n \log(d/d_0) + A_{mean} \quad (2)$$

where $RSSI_{mean}$ represents the mean RSSI. A_{mean} is the mean RSSI at distance d_0.

According to equation (2) the distance between two devices is then given by

$$d = d_0 10^{(Amean - RSSImean)/(10n)} \quad (3)$$

A. Optimization of Logarithmic Distance Path Loss Model

In order to apply the logarithmic distance path loss model to predict the distance between devices given the mean RSSI, suitable values of A, d_0 and n appearing in equation (3) were needed. A model function was created in python that returned the predicted distance given the mean RSSI, the values A, n and d_0 according to equation 3. As mentioned in Section III, various values of RSSI for measured actual distances were recorded. Some of these values and associated measured actual distances served as a data set to fit the model as shown in equation 3.

A built-in function called curve_fit in the python scipy module was used to fit the model appearing in equation (3). The curve_fit function uses nonlinear squares to fit the model to the data. The inputs to the curve_fit function were the model function as described before, the mean RSSI data, and the corresponding measured actual distances and the parameters d_0, A and n. The curve_fit function then returned the optimized values of n, A and d_0. The optimized values where latter used in the model to predict the distance between two devices.

V. Decision Tree

Decision Trees are a popular supervised learning method that can be applied in classification and regression problems [38]. They are capable of learning decision rules from the training dataset. These rules are usually expressed in form of if-then statements. When used in regression the decision tree model is piecewise smooth. According to [38], a Decision Tree can be set to have a certain maximum depth. However, as the depth increases, the tree rules tend to be complicated and such a tree is prone to overfitting.

In this work a Decision Tree was trained using various RSSI values in order to predict the distance between devices. The scikit-learn library [38] was used to implement the Decision Tree. The original dataset was first split into training and test sets. The training set comprised 80% of the original data whereas the test set comprised 20%. The training features were the RSSI values whereas the training labels were the actual distance between devices. A fit was then done on the training dataset. The fit was undertaken for models with varying maximum depth. It was observed that fitting was poor for models with small maximum depth, typically less than 5. A model with a maximum depth of 5 was therefore chosen. This model was then used to make predictions on the test dataset. It was observed that the larger the maximum depth model, the better the model fits the training dataset. However, models with larger maximum depths could not perform well when applied on the test set.

VI. Deep Neural Network Model

Deep learning is a subset of machine learning that is applied in several fields [39]. As pointed out in [39], in deep learning an artificial neural network consisting of multiple layers is used to model a problem. Among the layers are an input layer, a number of hidden layers and an output layer [39].

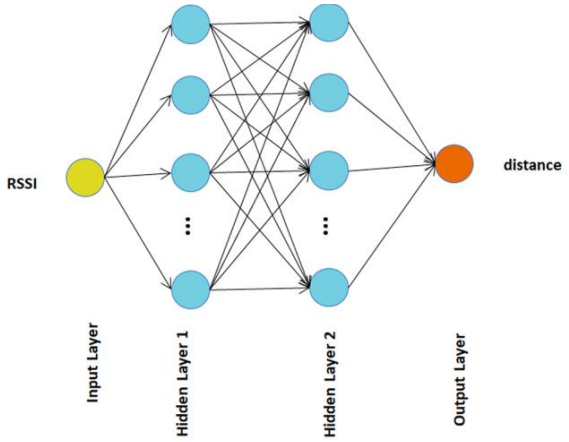


Fig. 2. Deep Neural Network Architecture for Predicting Distance between Two Devices.

In this work we propose a deep neural network (DNN) model to determine the distance between two devices using RSSI levels. Fig. 2 shows the deep neural network architecture for determining the distance between two devices. The input layer comprised one artificial neuron representing the RSSI signal which was subsequently normalized. For normalization, two quantities were first computed from the dataset. These are the mean and standard deviation. The normalized RSSI values were then computed by subtracting the mean from the original values and dividing the obtained result by the standard deviation. Two fully connected hidden layers were used in this study. The output layer predicted the distance between two devices. The programming language used to implement the model was python. The particular library used was tensorflow [40] and keras [41] as the application programming interface. The Rectified Linear Unit was used as the activation function. The Adaptive Moment Estimation (Adam) optimizer was chosen for this work. The learning rate was set to 0.01 as this was found to be appropriate. The loss function considered was the mean squared error. The original dataset for each platform was split into training and test sets where the training set comprised 80 percent of the original dataset. During training using keras, 10% of the training data was used for validation and the number of epochs was set to 200.

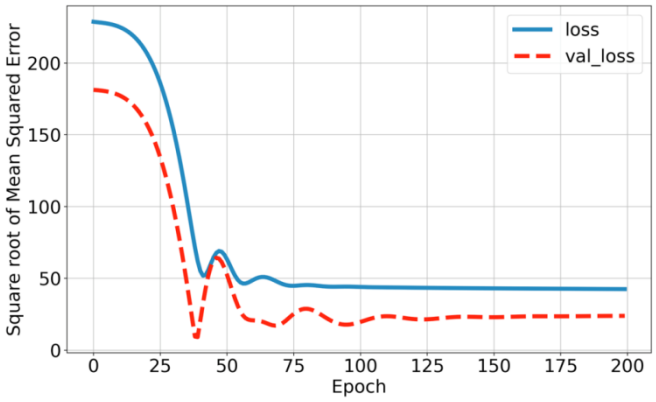


Fig. 3. Dependence of Square Root of Mean Squared Error on Epochs on Training Set (Loss) and Cross-Validation Sets (val_loss).

Fig. 3 shows the dependence of square root of mean squared error loss function on epochs. The loss functions on training set (loss) and cross-validation sets (val_loss) are shown in the Fig. 3.

The deep neural network model for each operating system was later converted to a model to be incorporated in a mobile application. For the iOS operating system, the model was converted to a coreml model using coremltools [42]. For the Android operating system, the model was converted to a tensorflow lite model according to [43]. As reported in [35], the RSSI is manufacturer dependent, therefore even for the same operating system, it varies from a device from one manufacturer to another.

VII. Comparison of Deep Neural Network, Decision Tree and Logarithmic Distance Path Loss Models

The accuracy of the deep neural network (DNN) model predictions was compared with the decision tree and logarithmic distance path loss model (LDPL). The root mean square deviation (RMSD), computed as the square root of the mean squared error (MSE) was used for the comparison. Equation 4 shows how the MSE is computed.

$$MSE = (1/(N-1))\ \Sigma_i\ (y_{predicted,i} - y_{true,i})^2 \quad (4)$$

In equation 4, $y_{predicted}$ represents the estimated target values, y_{true} represents the ground truth (correct) target values and N is the number of elements in the population. The MSE was determined using the mean_squared_error builtin function in a python sklearn.metrics module [44].

The root mean square deviation (RMSD) was then obtained according to equation 5.

$$RMSD = (MSE)^{1/2} \quad (5)$$

Fig. 4 shows the predictions of the deep neural network, decision tree as well as logarithmic distance path loss models for the iOS device. The maximum depth of the decision tree was set to 5. Also plotted is the training dataset. The trained models appear to reasonably fit the training dataset.

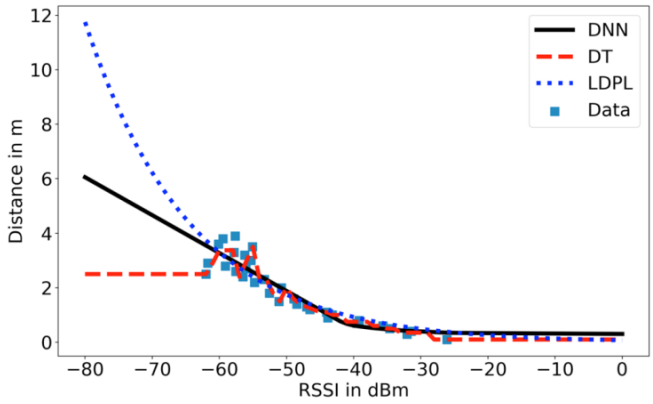


Fig. 4. Comparison of Actual Measured Data against Deep Neural Network (DNN), Decision tree (DT) and Logarithmic Distance Path loss (LDPL) Models Predictions during Training for the iOS Device.

Table I reports the RMSD computed according to equation (5) during training and testing of models. As reported in Table I, the DNN model performed better than the LDPL model during training and testing. However, the DT performed better than the DNN during training but the DNN performed better than DT during testing. This is attributed to overfitting by DT during training. Now, according to [28], it was reported that the DT performed better that the simplistic distance path loss model. This is also in agreement with the results reported in Table I.

TABLE I. Root Mean Square Deviation (RMSD) Calculation for Training and Testing the Deep Neural Network (DNN), Decision Tree (DT) and Logarithmic Distance Path Loss (LDPL) Models

RMSD on training the models in m			**RMSD on testing the models in m**		
DNN	***DT***	***LDPL***	***DNN***	***DT***	***LDPL***
0.41	0.15	0.46	0.44	0.49	0.57

The total processor (CPU) time to make model predictions on the entire training and test datasets used in Table I was

compared for the three models. Table II reports the total CPU time taken by the models in seconds.

TABLE II. TOTAL CPU TIME TAKEN TO MAKE MODEL PREDICTIONS ON THE DATASETS USED IN TABLE I.

CPU time taken during training in s			**CPU time taken during testing in s**		
DNN	***DT***	***LDPL***	***DNN***	***DT***	***LDPL***
0.091	0.00084	0.0008	0.095	0.00047	0.00098

According to Table II, the DT model took the least total CPU time to make predictions on the test dataset compared to LDPL and DNN models. However, as reported in Table I, the DNN model performed better than the DT and LDPL models in terms of predictions on the test dataset.

Owing to the fluctuation nature of the RSSI, a more accurate model for predictions is preferred despite the tradeoff in the CPU time taken to make predictions. As shown in Table II, the reported CPU times are all at sub-second level.

The processing power of mobile devices is generally lower than those of conventional desktop machines. It is worth mentioning that the DNN models have been further optimized to efficiently run on mobile devices as reported in [42] for iOS devices and [43] for Android devices. The DNN model was therefore adopted for prediction of distance between two devices.

VIII. DEVELOPMENT OF CROSS-PLATFORM CONTACT-TRACING MOBILE APPLICATION

In this work, a cross platform contact-tracing mobile application was developed. The targeted operating systems were iOS and Android. The application was developed in C# using Xamarin, a free, open source, cross-platform for building applications targeted at iOS and Android operating systems among others [45].

A number of features to be incorporated in the contact-tracing application were identified. These include the ability to detect the presence of another person running the same application on their device within an accepted range and notifying the users. The capability to notify a user if they have been exposed to a reported positive COVID-19 case in the past fourteen days was also included. In order to achieve this, the system used an already existing repository of COVID-19 patient data that contained unique diagnosis identifiers for patients. Furthermore, in case the user tested positive for COVID-19, the system was expected to provide an option to share their diagnosis identifier. This was an important feature that the application uses to notify others that they have come into contact with a positive case. A self-service feature where the user could query the application whether they have been in contact with a positive case was also included.

In order to meet these requirements, suitable technologies were identified. Bluetooth Low Energy (BLE) [46] was ideal for detecting when one user was closer to another. The Bluetooth LE Received Signal Strength Indicator (RSSI) levels described earlier were used to determine whether users were close to each other, within 2 metres for a period exceeding 15 minutes [18]. The deep neural network models described in Section VI were used to predict the distance between two users. The Global Positioning System (GPS) [47] / Cell Tower Triangulation [48] were used to determine the location in form of Latitude and Longitude coordinates. The location information was used for predicting COVID-19 hotspots.

To facilitate the contact tracing process, each mobile application user was assigned a unique random identifier that was later securely shared with another user who is in close proximity as determined by the deep neural network model. The Bluetooth Low Energy Generic Attribute Profile (GATT) framework was adopted in this case since it enables exchange of data between two devices [49]. According to GATT, the data is encapsulated in services where each service contains one or more characteristics.

The mobile application had GATT client and server capabilities. As a GATT server, the application was able to advertise services. To distinguish a mobile device running the contact tracing application from others, a unique service identifier following a universally unique identifier (UUID) format was generated for each device. Encapsulated in this service was a characteristic whose identifier was in universally unique identifier (UUID) format. The characteristic identifier was unique to the application and was used for contact tracing purposes. As a GATT client, the application was able to scan for a unique service associated with the contact tracing application and read the associated characteristics.

To implement the GATT capabilities in the application, a cross platform framework named Shiny was adopted [50]. Shiny supports BLE client and hosting among several features. This framework was also found to be convenient in that in supports backgrounding which allows an application to continue running even when sent to the background.

Fig. 5 is a top-level algorithm used to scan for BLE services, characteristics, RSSI and sending device manufacturer data among others. As shown in Fig. 5, the mobile application keeps scanning for devices and scan results are kept in a list. For every element of the list, the RSSI, manufacturer data as well as whether the device is connectable are obtained. With the obtained RSSI and sending device manufacturer data, an appropriate deep neural network (DNN) model is invoked to predict distance. In case the device is connectable, a connection was made to the device to discover offered services. The application checked for a particular service unique for contact tracing. If this service was available, the associated characteristics were obtained. The characteristics are used for contact tracing.

```
while scanning for devices:
    store scan results in a list
    for each result in a list:
        obtain RSSI, manufacturer data, IsConnectable value, and others from advertisement
        call appropriate DNN model to predict distance
        if IsConnectable value is true:
            connect to the device
            obtain discovered services
            if serviceUUID matches application UUID:
                obtain characteristics associated with service
                store RSSI, characteristicUUID, distance, manufacturer data, contact time for further processing
            end if
        end if
    end for
end while
```

Fig. 5. Top-Level Algorithm for Scanning for BLE Services, Characteristics and RSSI.

The predicted distance was used in further processing such as alerting the user in case the devices were too close. Datetime information was also captured in addition to Latitude/ Longitude coordinates. This information, together with the predicted distance from RSSI, sending device manufacturer data, characteristic UUID and contact time was stored in a local database in the user's mobile device.

After collecting this information, the contact tracing mobile application then securely transmitted this information to a central database of COVID-19 cases. An application programming interface (API) using DotNet core WebAPI development framework was written that allows for data to be shared from the mobile application to the database server using JavaScript Object Notation (JSON) as an exchange data format and HyperText Transfer Protocol (HTTP) protocol as a transport medium.

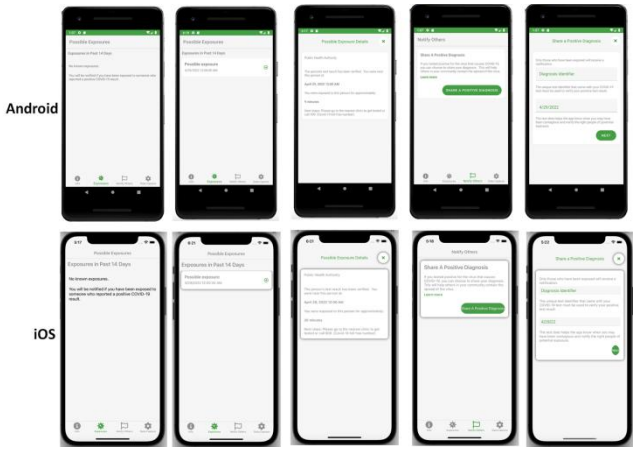


Fig. 6. Some Features of the Contact Tracing Mobile Application as Shown in Android (Top Row) and iOS (Bottom Row) Operating Systems.

A feature was also available that enables a user to determine in real-time whether they have been in contact with a positive COVID-19 case. The stored local contact tracing details for the last fourteen (14) days were then securely sent to the server and a Covid-19 database queried. If one of the contacts was recorded as a positive case in the COVID-19 database, the user was then alerted without revealing further details. To support this feature, the COVID-19 server offered a web service that was accessed through relational state transfer (REST) application programming interfaces (API).

Fig. 6 illustrates some features of the contact-tracing application. The mobile application provides real-time exposure alerts. The application also had a manual contact tracing feature to allow for positive diagnosed individuals to manually input the phone numbers of the contacts they have had met. In order to avoid abuse, a phone number verification feature was added. The application also incorporated location based services to get encrypted coordinates to be used in prediction of epicentres.

Furthermore, the Latitude, Longitude and date time information from contact-tracing applications was sent securely to the Ministry of Health databases. This data can be displayed in real-time in a map for identification of possible hotspots as shown in Fig. 7.

In this section it has been shown how the GATT framework was used to encapsulate and exchange data between Bluetooth devices embedded with GATT framework. Furthermore, the implementation of the deep neural network model in order to predict distance between communicating devices has been reported.

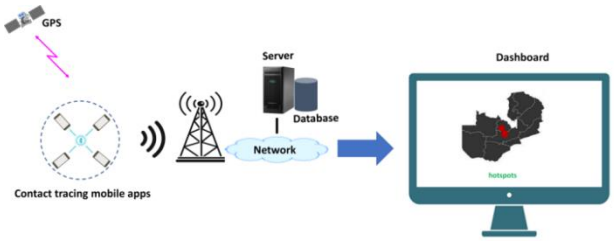


Fig. 7. Using Location and Datetime Information from Contact-Tracing Applications for Identification of Hotspots.

IX. Conclusion

A number of digital contact tracing applications have been applied world-over to facilitate the contact-tracing process. In this work, a cross platform contact tracing application that uses deep neural network models and Bluetooth Low Energy Generic Attribute Profile framework to determine and inform users of exposure to COVID-19 has been developed. The performance of deep neural network models has been evaluated against other models. The reported results show that the deep learning model performs well during testing.

The proposed deep learning model appears to learn the nonlinear relationship between distance and RSSI values better compared to analytic and decision tree models. The analytic model had four parameters according to equation 3. This suggests that the analytic model overlooked some parameters that are useful in determining the distance between communicating devices. As regards to decision trees, a major limitation is their likelihood to overfit the data on the training set as the maximum depth increases. The smaller the maximum depth, the less is the generalizing ability. On the other hand the larger the maximum depth, the larger the likelihood of overfitting on the training set resulting in less accuracy on the test set.

The developed contact-tracing application can be beneficial not only to COVID-19 prediction but also to other pandemics.

References

[1] D. Cucinotta, M. Vanelli, "WHO declares COVID-19 a pandemic," Acta Bio Medica: Atenei Parmensis, vol 91, issue 1, pp. 157 –160, 2020

[2] S. Basu, "Effective contact tracing for COVID-19 using mobile phones: an ethical analysis of the mandatory use of the aarogya setu application in India," Cambridge Quarterly of Healthcare Ethics, vol 30, issue 2, pp. 262 – 271, 2021

[3] K.T. Eames, M.J. Keeling, "Contact tracing and disease control," Proceedings of the Royal Society of London. Series B: Biological Sciences, vol 270, issue 1533, pp. 2565 –2571, 2003

[4] J. Müller, M. Kretzschmar, "Contact tracing-Old models and new challenges," Infectious Disease Modelling, vol 6, pp. 222 –231, 2021

[5] D. Klinkenberg, C. Fraser, H. Heesterbeek, "The effectiveness of contact tracing in emerging epidemics," PloS one, vol 1, issue 1, p. e12, 2006

[6] L.O. Danquah, N. Hasham, M. MacFarlane, F.E. Conteh, F. Momoh, A.A. Tedesco, A. Jambai, D.A. Ross, H.A. Weiss, "Use of a mobile application for Ebola contact tracing and monitoring in northern Sierra Leone: a proof-of-concept study," BMC infectious diseases, vol 19, issue 1, pp. 1 –12, 2019

[7] J.A. Sacks, E. Zehe, C. Redick, A. Bah, K. Cowger, M. Camara, A. Diallo, A.N.I. Gigo, R.S. Dhillon, A. Liu, "Introduction of mobile health tools to support Ebola surveillance and contact tracing in Guinea," Global Health: Science and Practice, vol 3, issue 4, pp. 646 – 659, 2015

[8] K.C. Swanson, C. Altare, C.S. Wesseh, T. Nyenswah, T. Ahmed, N. Eyal, E.L. Hamblion, J. Lessler, D.H. Peters, M. Altmann, "Contact tracing performance during the Ebola epidemic in Liberia, 2014-2015," PLoS neglected tropical diseases, vol 12, issue 9, p.e0006762, 2018

[9] G. Mosweunyane, T. Seipone, T.Z. Nkgau, O.J. Makhura, " Design of a USSD system for TB contact tracing," In IASTED International Conference Health Informatics (AfricaHI 2014), ACTAPRESS, 2014

[10] M. Shahroz, F. Ahmad, M.S. Younis, N. Ahmad, M.N.K. Boulos, R. Vinuesa, J. Qadir, "COVID-19 digital contact tracing applications and techniques: A review post initial deployments," Transportation Engineering, vol 5, p. 100072, 2021

[11] M. Nazayer, S. Madanian, F. Mirza, "Contact-tracing applications: a review of technologies," BMJ Innovations, vol 7, issue 2, pp. 368 – 378 2021

[12] TraceTogether, https://www.tracetogether.gov.sg, date accessed 03 October 2021

[13] J. Bay, J. Kek, A. Tan, C.S. Hau, L. Yongquan, J. Tan, T.A. Quy, " BlueTrace: A privacy-preserving protocol for community-driven contact tracing across borders," Government Technology Agency-Singapore, Tech. Rep, 18, 2020

[14] C. Wymant, L. Ferretti, D. Tsallis, M. Charalambides, L. Abeler-Dörner, D. Bonsall, R. Hinch, M. Kendall, L. Milsom, M. Ayres, C. Holmes, M. Briers, C. Fraser, "The epidemiological impact of the NHS COVID-19 App, " preprint at go.nature.com/2m4scfk, 2021

[15] Italy: Government Implements Voluntary Contact Tracing App to Fight COVID-19. [Web Page] Retrieved from the Library of Congress, https://www.loc.gov/item/global-legal-monitor/2020-11-09/italy-government-implements-voluntary-contact-tracing-app-to-fight-covid-19/.

[16] Spain: Government's Contact Tracing App Now in Operation Throughout Country. [Web Page] Retrieved from the Library of Congress, https://www.loc.gov/item/global-legal-monitor/2020-11-20/spain-governments-contact-tracing-app-now-in-operation-throughout-country/.

[17] Privacy-Preserving Contact Tracing - Apple and Google, https://www.apple.com/covid19/contacttracing, date accessed 03 October 2021.

[18] L. Dyani, "Contact-Tracing Apps Help to Reduce Covid Infections," Nature, vol 591, pp. 18 –19, 2021

[19] A. Narzullaev, Z. Muminov, M. Narzullaev, "Contact Tracing of Infectious Diseases Using Wi-Fi Signals and Machine Learning Classification," In 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), pp. 1 – 5, 2020

[20] A. Roy, F.H. Kumbhar, H.S. Dhillon, N. Saxena, S.Y. Shin, S. Singh, "Efficient monitoring and contact tracing for COVID-19: A smart IoT-based framework," IEEE Internet of Things Magazine, vol 3, issue 3, pp. 17 – 23, 2020

[21] Y. Sahraoui, L. De Lucia, A.M. Vegni, C.A. Kerrache, M. Amadeo, A. Korichi, "TraceMe: Real-Time Contact Tracing and Early Prevention of COVID-19 based on Online Social Networks," In 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC), pp. 893 – 896, 2022

[22] K. Carteri, G. Berman, M. Garcia-Herranz, V. Sekara, "Digital contact tracing and surveillance during COVID-19 General and Child-specific Ethical Issues," https://www.unicef-irc.org/publications/pdf/IRB2020-11.pdf, date accessed 9 October 2021

[23] T. Jiang, Y. Zhang, M. Zhang, T. Yu, Y. Chen, C. Lu, J. Zhang, Z. Li, J. Gao, S. Zhou, "A survey on contact tracing: the latest advancements and challenges," ACM Transactions on Spatial Algorithms and Systems (TSAS), vol 8, issue 2, pp.1 – 35, 2022

[24] O. Megnin-Viggars, P. Carter, G.J. Melendez-Torres, D. Weston, G.J. Rubin, "Facilitators and barriers to engagement with contact tracing during infectious disease outbreaks: A rapid review of the evidence," PloS one, vol 15, issue 10, p. e0241473, 2020

[25] E. Bandara, X. Liang, P. Foytik, S. Shetty, C. Hall, D. Bowden, N. Ranasinghe, K. De Zoysa, "A blockchain empowered and privacy preserving digital contact tracing platform," Information Processing & Management, vol 58, issue 4, p.102572, 2021

[26] M. Torky, E. Goda, V. Snasel, A.E. Hassanien, "COVID-19 Contact Tracing and Detection-Based on Blockchain Technology," In Informatics, vol. 8, issue 4, p. 72, Multidisciplinary Digital Publishing Institute, 2021

[27] H. Xu, L. Zhang, O. Onireti, Y. Fang, W.J. Buchanan, M.A. Imran, "BeepTrace: blockchain-enabled privacy-preserving contact tracing for COVID-19 pandemic and beyond," IEEE Internet of Things Journal, vol 8, issue 5, pp. 3915 – 3929, 2020

[28] P.C. Ng, P. Spachos, K.N. Plataniotis, "COVID-19 and your smartphone: BLE-based smart contact tracing," IEEE Systems Journal, vol 15, issue 4, pp. 5367 – 5378, 2021

[29] P.C. Ng, P. Spachos, S. Gregori, K.N. Plataniotis, "Epidemic Exposure Tracking With Wearables: A Machine Learning Approach to Contact Tracing," IEEE Access, vol 10, pp. 14134 –14148, 2022

[30] L. Bai, F. Ciravegna, R. Bond, M. Mulvenna, "A low cost indoor positioning system using bluetooth low energy," IEEE Access, vol 8, pp. 136858 – 136871, 2020

[31] J. Paek, J. Ko, H. Shin, "A measurement study of BLE iBeacon and geometric adjustment scheme for indoor location-based mobile applications," Mobile Information Systems, 2016

[32] H. Meijerink, C. Mauroy, M.K. Johansen, S.M. Braaten, C.U.S. Lunde, T.M. Arnesen, E.H. Madslien, "The first GAEN-based COVID-19 contact tracing app in Norway identifies 80% of close contacts in "real life" scenarios, " Frontiers in digital health, vol 3, 2021

[33] J. Kalezhi, M. Chibuluma, C. Chembe, V. Chama, F. Lungo, D. Kunda, "Modelling Covid-19 infections in Zambia using data mining techniques," Results in Engineering, vol 13, p. 100363, 2022

[34] Zambia National Public Health Institute, Available: https://znphi.co.zm, date accessed: 19 February 2021

[35] Proximity and RSSI, https://www.bluetooth.com/blog/proximity-and-rssi/ Date accessed: 16 March 2022

[36] L. Flueratoru, V. Shubina, D. Niculescu, E.S. Lohan, "On the High Fluctuations of Received Signal Strength Measurements with BLE Signals for Contact Tracing and Proximity Detection," IEEE Sensors Journal, 2021

[37] G. Li, E. Geng, Z. Ye, Y. Xu, J. Lin, Y. Pang, "Indoor positioning algorithm based on the improved RSSI distance model," Sensors, vol 18, issue 9, p. 2820, 2018

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, "Scikit-learn: Machine learning in Python," the Journal of machine Learning research, vol 12, pp. 2825 –2830, 2011

[39] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning, " Nature, vol 521, pp. 436 – 444, 2015

[40] M. Abadi et al, "TensorFlow: Large-scale machine learning on heterogeneous systems," Software available from tensorflow.org, 2015

[41] F. Chollet et al, "Keras," Available at: https://github.com/fchollet/keras, 2015

[42] TensorFlow 2 Conversion, https://coremltools.readme.io/docs/tensorflow-2 Date accessed: 18 March 2022

[43] tf.lite.TFLiteConverter : TensorFlow Core v2.8.0, https://www.tensorflow.org/api_docs/python/tf/lite/TFLiteConverter Date accessed: 18 March 2022

[44] G. Van Rossum, F.L. Drake, "Python 3 Reference Manual," Scotts Valley, CA: CreateSpace, 2009

[45] Xamarin: Open-source mobile app platform for .NET , https://dotnet.microsoft.com/apps/xamarin, date accessed: 8 October 2021

[46] Bluetooth Technology Overview, https://www.bluetooth.com/learn-about-bluetooth/tech-overview/, date accessed: 8 October 2021

[47] Satellite Navigation - Global Positioning System (GPS), https://www.faa.gov/about/office_org/headquarters_offices/ato/service_units/techops/navservices/gnss/gps/, date accessed: 8 October 2021

[48] J. Yang, A. Varshavsky, H. Liu, Y. Chen, M. Gruteser, "Accuracy characterization of cell tower localization," In Proceedings of the 12th ACM international conference on Ubiquitous computing 2010 Sep 26, pp. 223 –226, 2010

[49] K. Townsend, C. Cufí, R. Davidson, "Getting started with Bluetooth low energy: tools and techniques for low-power networking, " O'Reilly Media, Inc., 2014

[50] shinyorg/shiny: A Xamarin Framework for Backgrounding & Device Hardware Services, https://github.com/shinyorg/shiny Date accessed: 18 March, 2022.

A Hybrid 1D-CNN-Bi-LSTM based Model with Spatial Dropout for Multiple Fault Diagnosis of Roller Bearing

Gangavva Choudakkanavar, J. Alamelu Mangai
Department of Computer Science and Engineering
Presidency University, Bangalore, India

***Abstract*—Fault diagnosis of roller bearings is a crucial and challenging task to ensure the smooth functioning of modern industrial machinery under varying load conditions. Traditional fault diagnosis methods involve preprocessing of the vibration signals and manual feature extraction. This requires domain expertise and experience in extracting relevant features to accurately detect the fault. Hence, it is of great significance to implement an intelligent fault diagnosis method that involves appropriate automatic feature learning and fault identification. Recent research has shown that deep learning is an effective technique for fault diagnosis. In this paper, a hybrid model based on 1D-CNN (One-Dimensional Convolution Neural Networks) with Bi-LSTM (Bi-directional Long-Short Term Memory) is proposed to classify 12 different fault types. Firstly, vibration signals are given as input to 1D-CNN to extract intrinsic features from the input signals. Then, the extracted features are fed into a Bi-LSTM model to identify the faults. The performance of the proposed method is enhanced by applying Softsign activation function in the Bi-LSTM layer and Spatial Dropout in the neural network. To analyze the effectiveness of the proposed method, Case Western Reserve University (CWRU) bearing data is considered for experimentation. The results demonstrated that the proposed model has attained an accuracy of 99.84% in classifying the various faults. The superiority of the proposed method is verified by comparing the predictive accuracy of the proposed method with the existing fault diagnosis methods.**

***Keywords*—Fault diagnosis; roller bearing; deep learning; 1D-CNN; Bi-LSTM; spatial dropout**

I. Introduction

Roller Bearing (RB) is a key component of any rotating machinery where rotation is involved. It is widely used in various industries such as transportation, agriculture, aerospace, medical domain and so on. RB is more susceptible to damage due to its continuous rotation with varying loads and pressure. Due to which there's a break-down of the entire machine which results in magnificent economic loss and severe safety accidents [1]. Therefore, it is very much essential to diagnose the roller bearing fault accurately because each fault type exhibits distinct characteristics and the fault may exist in any of the components such as Inner Race (IR), Outer Race (OR) and Ball.

Traditional vibration-based bearing fault diagnosis methods involved mainly three steps as data pre-processing, feature extraction, and fault classification. The vibration signals collected from sensors represents the information about bearing condition. In order to classify and detect the faults, many signal processing techniques have been discussed through analysis of signal characteristics in various domains such as time, frequency and time-frequency domain [2]. Due to the non-stationary nature of vibration signal, various feature extraction techniques such as Short-Time Fourier Transform (STFT), Wavelet Analysis (WA), Empirical Mode Decomposition (EMD), etc. were applied to extract the features [3]. Once the features are extracted and selected then those features are fed into the network model for classification.

Recently, Deep Learning (DL) technology has gained more importance in various domains such as image processing, natural language processing, speech recognition and so on. It uses multiple layers of the network to learn and extract relevant features from raw data and identifies the pattern for classification or recognition problems. Roller bearing's vibration data has similar dimensionality as that of image or speech. Hence, DL architecture can be used to diagnose roller bearing fault by transforming vibration signal into the framework of pattern recognition problem. DL model has an ability in automatic feature learning and classification that involves automatic feature extraction and identification of the faults accurately [4-5].

In this research, a hybrid method based on 1D-CNN-Bi-LSTM with Spatial Dropout is proposed for multiple fault diagnosis of roller bearing. Initially, one-dimensional raw vibration signal is collected and input into CNN model. Then, CNN extracts feature information from the signals and these extracted features are provided to Bi-LSTM network model to acquire the failure information to identify 12 types of bearing faults. For experimentation, CWRU dataset is being used to analyze the effectiveness of the method.

The rest of this paper is organized as follows: In Section II related work is discussed; Section III describes proposed methodology architecture which includes one-dimensional CNN and Bi-LSTM models. Section IV illustrates an experimental setup of bearing data collection; and Section V shows the discussion of results and its analysis.

II. Related Work

Many researchers have applied various deep learning models for fault diagnosis such as Deep Neural Networks [6], Long Short-Term Memory [7], Deep Belief Networks [8], Deep Auto-encoders [9], Gated Recurrent Unit Networks [10],

and Convolutional Neural Networks [11] and so on. Among these, CNN has received more importance in the study of roller bearing defect diagnosis. Abed et al. [12] proposed a robust approach for fault diagnosis of Brushless DC Motors through feature extraction and reduction using discrete wavelet transform (DWT) and orthogonal fuzzy neighborhood discriminant analysis (OFNDA) from vibration and current signals and RNN model was used for classification of faults. P. Zou et al. [13] focused on empirical mode decomposition (EMD) method which was combined with LSTM to obtain kurtosis value by extracting intrinsic mode functions (IMF) components and long-term dependencies from vibration signals to monitor the health status of an electrical machine. Cao, Lixiao et al. [14] constructed a fault diagnosis framework by extracting ten time-domain statistical features from vibration signals under varying load conditions and these features were fed into deep Bi-directional LSTM to identify the faults of Wind Turbine Gearbox. In [15], Mel Frequency Cepstral Coefficient features are obtained from vibration signals and given these features as input to Random Forest and eXtreme Gradient Boosting algorithms for diagnosis of roller bearing fault.

Zheng Wang et al. [16] discussed an architecture to obtain unsupervised H-statistic value from sensor time-series data based on deep LSTM and CNN for performance degradation valuation of roller bearing. Shichao and Haibin proposed a bearing fault diagnosis model in which 1D-CNN with LSTM is implemented, which adaptively extracted potential features from the original vibration signal and ensured the validity of the features through merging of pooling layers of max and average values to down sample the features. Then, LSTM was employed to acquire the dependencies among features of time-domain signals to perform fault classification [17]. Zhe Yuan et al. [18] presented a fault recognition approach for roller bearing using Multiscale CNN and Gated Recurrent Unit Network (GRUN) by providing multiple time scaled vibration data into the CNN to train the model and added the gated recurrent unit network to make the model predictive with an attention mechanism. In [19], the proposed adaptive anti-noise neural network architecture employed random sampling approach and boosted CNN with the exponential linear activation function to enhance the adaptability of the network without manual feature selection. GRUN was implemented to learn the features processed by CNN and classify the faults. This approach solved the problem of bearing fault diagnosis under changing load conditions and heavy noise.

Wenbing Yu et al. [20] discussed an intelligent fault diagnosis method for identifying ten different bearing faults based on lightweight MobileNet CNN by considering Western Reserve University dataset for evaluating the model and also computed average precision, recall and F1 score which resulted into 96%, 82% and 88%, respectively. Kai Gu et al. [21] discussed a novel diagnostic method to accurately identify the fault status of bearing based on LSTM and DWT for multi-sensors by obtaining fault details in both frequency and time scales through DWT and LSTM algorithm was used to characterize the long-term dependency information hidden in the time series data of a signal. In [22], a combined wavelet regional correlation threshold denoising (WRCTD) algorithm with CNN-LSTM was proposed for fault detection. WRCTD algorithm utilized the regional association of the wavelet decomposition coefficients and 3σ criterion to reduce noise in the raw sensor data and CNN-LSTM model reduced the hidden features of the pre-processed signal data to identify the fault type of the harmonic reducer under multiple working conditions. A novel fault diagnosis method was presented through application of sliding window processing to integrate the feature and time delay information from multivariate time series samples and then, the samples obtained were fed into the CNN-LSTM model to perform feature learning and capture time delay information to diagnose the fault of Tennessee Eastman chemical process [23]. A robust approach was proposed in [24] to predict the Remaining Useful Life (RUL) of roller bearing with combination of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and LSTM to detect the damage state and identify the abnormal state of bearing to estimate the RUL through feature extraction from signals. A new convolution-based bidirectional long and short-term memory network method was proposed to predict RUL, in which CNN was used to obtain feature information and BI-LSTM to acquire time-frequency information from the signals to construct health indicators (HI) and the experiments conducted on the PRONOSTIA bearing dataset showed that the proposed method performed better compared to other methods [25].

This work uses deep learning technique for fault diagnosis and it's motivated by the fact that deep learning involve automatic feature extraction whereas Machine learning needs manual feature extraction, in which prior domain knowledge and expertise is required.

III. Proposed Methodology

In this research, a multi-class fault diagnosis method is proposed based on 1D-CNN with Bi-LSTM to classify various faults. The advantage of CNN lies in automatic feature extraction and Bi-LSTM in handling gradient loss and explosion. The main goal is to diagnose the 12 different fault types using 1D-CNN with Bi-LSTM model by collecting vibration signals from CWRU dataset which contains set of ball bearings having localized faults. The proposed 1D-CNN-Bi-LSTM model consists of four convolutional and pooling layers, a Bi-LSTM layer, a LSTM layer and one fully connected layer as shown in Fig. 1.

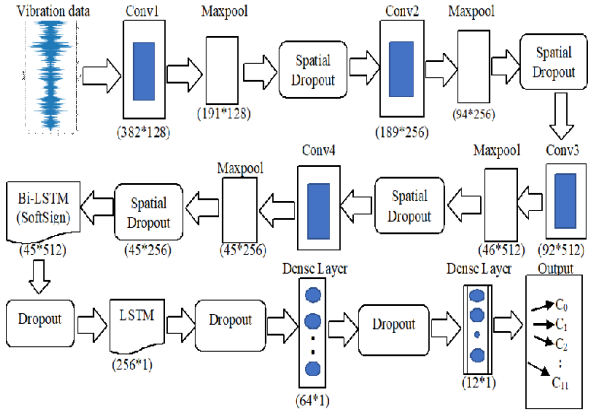


Fig. 1. Block Diagram of the Multi-class Fault Diagnosis Method.

Firstly, the raw input signals are input to the model, then convolution layers and pooling layers helps in automatic feature extraction. Next, these features are passed to Bi-LSTM and LSTM layers to highlight the features and finally, dense layer is used to classify the various faults. Bi-LSTM layer uses softsign activation function to improve the performance of the model. For verifying the effective performance of the proposed method, publicly available bearing dataset is considered [26]. The vibration data is measured for four operational conditions such as:

1) Normal Bearing- No fault, sampling frequency of 12kHz with 1797rpm (rotations/min).

2) Fault in Outer race - Sampling frequency of 12kHz with 1797rpm.

3) Fault in Inner race - Sampling frequency of 12kHz with

4) 1797rpm.

5) Ball fault- Sampling frequency of 12kHz with 1797rpm.

Details of each layer used in the proposed method is explained in the following subsections.

A. 1D-CNN

CNN is a deep learning algorithm which was originally proposed for processing of visual data. It is more effective in identifying image patterns in a stratified way from simple to complex features because of the two important properties such as weight sharing and spatial pooling. CNN consists of 3 layers namely convolutional layer, pooling layer, and fully connected layer. The convolution layer converts the input data into smaller feature maps through convolutional kernels by performing a summation of multiplications between the vectors of input data and weight coefficients [27]. In this paper, 1D-CNN is constructed, whose convolutional kernels and feature maps are all one-dimensional because of the one-dimensional characteristics of mechanical vibration signals.

Suppose 'x' is an input to 1D-CNN, then the output of the convolutional layer is computed as given in (1):

$$y_{i,\,j,\,k} = f\left(\sum_{i=1}^{m} x_{i,\,k} * w_{j,\,i} + b_i\right) \quad (1)$$

In equation (1), 'f' represents an activation function, which is typically a hyperbolic tangent, ReLu (Rectified Linear Unit), or sigmoid function; 'm' is number of samples ($1 \le i \le m$); 'p' is length of the convolutional kernels ($1 \le j \le p$); 'n' is length of the input data ($1 \le k \le n$); * represents convolution operation; $w_{j,i}$ is the weight and b_i is the bias.

The pooling layer is the sub-sampling layer to compress the size of feature maps. Down sampling is performed to minimize the dimensionality of the output from the previous convolution layer by moving the filter window from starting point to the end of feature map. Then a maximum or average of each part of the feature map is considered to represent each corresponding area. The role of pooling layer is to reduce the number of parameters and the computation in the network, so that it prevents overfitting and improves the generalization ability of the model. Max pooling is frequently used in the pooling layer which is computed as maximum of the previous feature maps. It is expressed as given in (2).

$$z_{i,\,j,\,k} = \max\left(x_{2i-1,\,j,\,k}, x_{2i,\,j,\,k}\right) \quad (2)$$

where, $1 \le l \le m/2$

Fully connected or Dense layer plays the role of classifier in CNN. For a multi-class classification problem, usually softmax is applied in the dense layer to ensure the range of output value lies between 0 and 1, and sum equals to 1. The predicted output represents the probability and value with the highest probability is considered as the final predicted result. The output of 1D-CNN is an input to the Bi-LSTM model to reduce variance in time series.

B. Batch Normalization

It is a regularization technique, which avoids model overfitting. In the training process of the deep neural networks, the distribution of inputs to the layers deep in the network which keeps changing for each mini batch as the weights are updated. This problem is known as "internal covariate shift". It delays the network to converge during the training phase. To avoid this problem, Batch normalization standardizes the input to each layer after every mini-batch and hence accelerates the network training. It is usually applied either before or after the activation functions of each hidden layer.

The process of batch normalization is shown in Fig. 2 [28]. It shifts the values of the input distribution to a hidden layer, such that the mean of these values is zero (zero centered) and then normalizes the inputs. It creates two parameter vectors for each layer, one with the scaled values and the second vector with the shifted values of the inputs to the layers.

The output scale vector 'γ' and the output offset vector 'β' are learnt through backpropagation. The final input mean vector 'μ' and the final input standard deviation vector 'σ' are estimated using exponential moving average during training.

C. Dropout

In a fully connected neural network, the probability of co-adaptation among the neurons is likely higher. As a result, the features extracted by the neurons for learning is more or less similar. This co-adaptation makes the model to overfit the training data and generalize poorly on unseen test data. Dropout is a technique to overcome this problem. It chooses a specified percentage of neurons randomly to be dropped during training by making their connection weights to zeros. Repeated application of this technique creates an ensemble of network architecture with a different set of neurons and their weights are dropped in each architecture as shown in Fig. 3(a) and (b) [29].

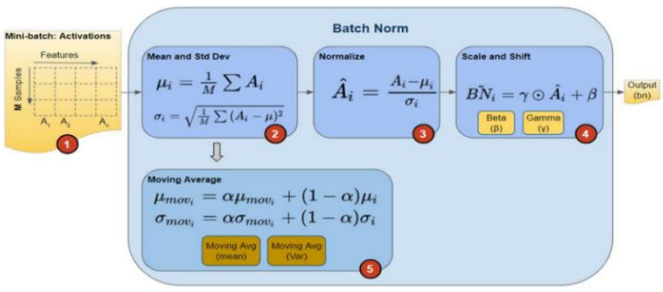
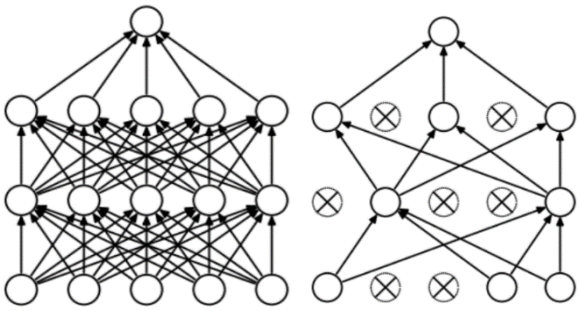


Fig. 2. Batch Normalization.



(a) (b)

Fig. 3. (a). Standard Neural Network. (b). Neural Networ with Dropouts.

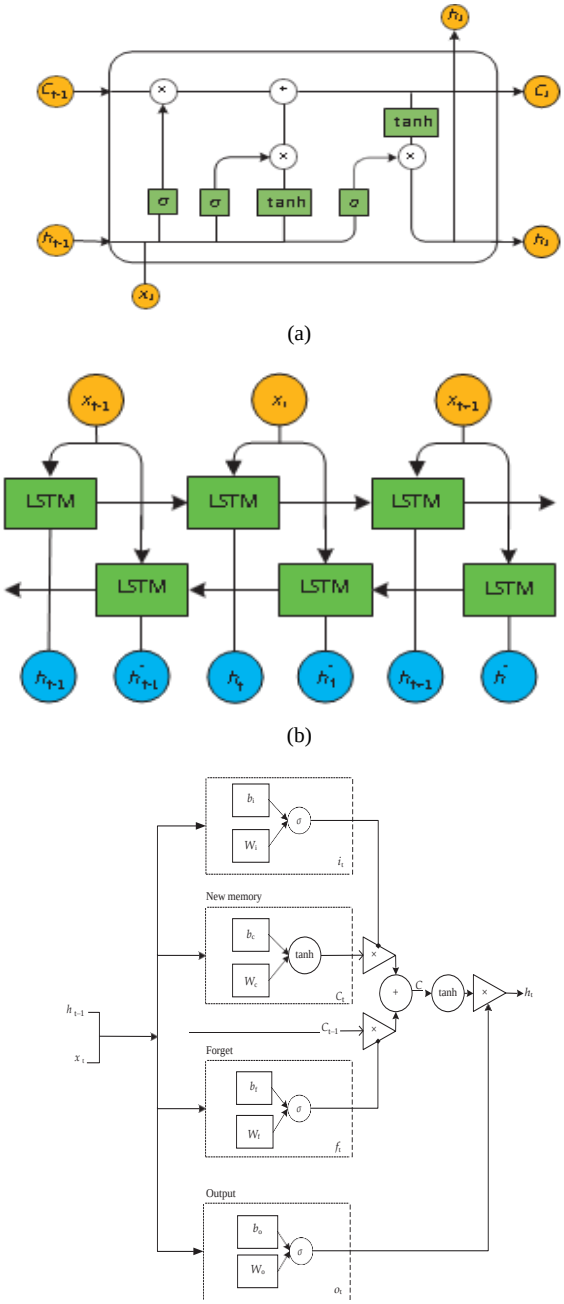
The weights of these dropped neurons are not updated during backpropagation. When some of the neurons are dropped, the other neurons take the responsibility of propagating the features to the subsequent layers of the network in the forward pass. Hence it prevents a sort of co-adaptation among the neurons and makes the network less reliable to the learning units, their weights and existence. All these factors help to generalize the model well on the test samples. Dropouts in convolutional layers are applied to the individual cells of the feature map/kernel and are called spatial dropouts. Dropouts applied to the hidden layers are regular dropouts.

D. Bi-LSTM (Bi-Directional Long-Short Term Memory)

Bi-Directional Long-Short Term Memory is a type of Recurrent Neural Network (RNN), which is a deep learning technique that is used to categorize and regress timeseries data such as audio, text forecasting and so on. Bi-LSTM combines LSTM layers from both directions. Hence, it captures long-term dependencies between signal patterns by making the flow of information in both forward and backward directions. There are 3 components in LSTM, namely i) forget gate, ii) update gate, and iii) output gate. The forget gate eliminates the irrelevant information which is received from the preceding unit. The update gate performs addition of information to the cell state, and the output gate selects the relevant information from the present cell state and gives the output [30]. The LSTM gating structure manages the information by enabling the memory cells to preserve long-term dependencies through selective passage. It avoids the problems of gradient loss and gradient explosion by strengthening the weight of relevant information and weakening the weight of irrelevant ones. The structure of the LSTM cell is shown in Fig. 4(a).

The LSTM network cannot make use of the full data while processing the time series signals because it processes the data only in one direction. Hence, the Bi-LSTM network is implemented, which contains LSTM layers overlaid on each other in reverse direction. It improves the performance by enabling the model to make efficient use of the main features. The unit structure of the Bi-LSTM network is shown in Fig. 4(b).

The internal processing of the LSTM cell is shown in Fig. 4(c).



(a)

(b)

(c)

Fig. 4. (a). LSTM Cell Structure, (b). Bi-LSTM Unit Structure, (c). The Internal Process of the LSTM Cell.

The inputs W_i, W_o, W_f, W_c represents weights and b_i, b_o, b_f, b_C represents bias vectors of input gate, output gate, forget gate and cell state respectively. The input of current state and output of the previous state is represented as x_t and h_{t-1}. The input value C_t' at moment 't' is calculated by applying the tanh activation function on the result obtained by computing the matrix product of vector $[h_{t-1}, x_t]$ with W_c and b_c as given in (3).

$$C_t^{'} = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \quad (3)$$

The parametric value for each gate i.e., f_t-forget gate, i_t-input gate, and o_t-output gate at moment 't' is calculated by applying the activation function as shown in (4), (5) and (6).

$$f_t = \sigma\left(W_f \cdot \left[h_t\text{-}1, x_t\right] + b_f\right) \quad (4)$$

$$i_t = \sigma\left(W_i \cdot \left[h_t\text{-}1, x_t\right] + b_i\right) \quad (5)$$

$$o_t = \sigma\left(W_o \cdot \left[h_t\text{-}1, x_t\right]\right) + bo \quad (6)$$

An element-wise product of '*ft'* with the last cell state 'C_{t-1}' determines the info that is to be forgotten and remembered by realizing the control on C_{t-1} and the element-wise product of 'i_t' with the current input cell state C_t' determines the info in C_t' that needs to be stored and used. The state value 'C_t' of the hidden node at time '*t'* is calculated as given in (7).

$$C_t = f_t * C_{t-1} + i_t * C_t' \quad (7)$$

The output value 'h_t' at time '*t'* is computed as product of tanh function applied on unit state C_t, and output gate o_t, as given in (8).

$$h_t = o_t * \tanh(C_t) \quad (8)$$

E. Softsign Activation Function

The softsign function squishes its input to a range of -1 to +1 as like tanh. The function and its derivative are defined as given in (9) and (10).

$$\text{softsign } f(x) = \frac{x}{1+|x|} \quad (9)$$

$$f'(x) = \frac{1}{(1+|x|)^2} \quad (10)$$

Unlike tanh, this function has a flatter curve, its derivative descends slowly, and is less saturated. Functions that are more saturated, have their gradients vanishing quickly before reaching the initial layers of the network during backpropagation [31]. Hence, softsign solves this vanishing gradient problem better than tanh. Softsign converges in polynomial time whereas tanh converges in exponential time. Since softsign transforms the inputs between -1 to +1, the negative values enable the LSTM gates to delete the information when required.

IV. Experimental Setup

As the benchmark study, CWRU bearing dataset has been widely considered by many researchers for condition monitoring and fault diagnosis. An experimental setup of CWRU is shown in Fig. 5. It consists of a 2- hp (horsepower) motor, encoder, torque transducer, dynamometer, electric motor and so on. The deep groove ball bearing was mounted on the drive end of the motor to support the shaft which needs to be tested. An accelerometer was positioned above the bearing base of the drive end to measure the vibration signals. The load considered was about 1HP with 1772 rpm, and sampling rate of 12 kHz. Fault was induced in each component of rolling bearing by electro discharge with varying diameters of 0.007, 0.014, and 0.021 inches (1 inch = 25.4 mm).

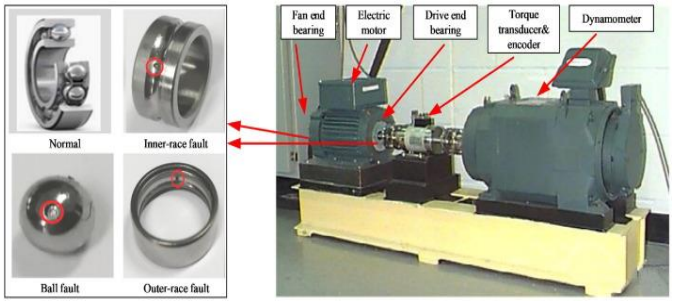


Fig. 5. Experimental Set-up of CWRU Data Collection.

In total, 12 bearing fault types with respect to BF, IR, OR and normal bearing were considered in this work as given in Table I. To label the fault types, One-Hot coding technique was used. In this experiment, to confirm adequate training size, 80% of the data was randomly chosen as training set and 20% as test set. For validation of the model, 10% of the training set was selected randomly to adjust model parameters.

V. Results and Discussion

In this proposed work, the vibration data is collected for 12 different bearing conditions that is provided by CWRU. The description of various fault types and count of samples considered for each fault class from the experimental setup is given in Table I.

Time-domain features for normal and faulty bearings are shown in Fig. 6(a)-(d).

The proposed hybrid model is implemented using the Python's deep learning modules i.e. Tensorflow and Keras [32].

TABLE I. Summary of 12 Drive End Fault Data of CWRU

Class	Types of Bearing Faults	No. of samples	Data description
C0	0.007-Ball	319	Ball fault level =0.007
C1	0.007-InnerRace	315	Inner Race fault level=0.007
C2	0.007-OuterRace12	318	Outer Race fault level at 12'o clock position = 0.007
C3	0.014-Ball	317	Ball fault level =0.014
C4	0.014-InnerRace	317	Inner Race fault level=0.014
C5	0.014-OuterRace6	317	Outer Race fault level at 6'o clock position = 0.014
C6	0.021-Ball	317	Ball fault level =0.021
C7	0.021-InnerRace	318	Inner Race fault level=0.021
C8	0.021-OuterRace12	317	Outer Race fault level at 12'o clock position = 0.021
C9	0.028-Ball	314	Ball fault level =0.028
C10	0.028-InnerRace	314	Inner Race fault level=0.028
C11	Normal	634	Normal Bearing

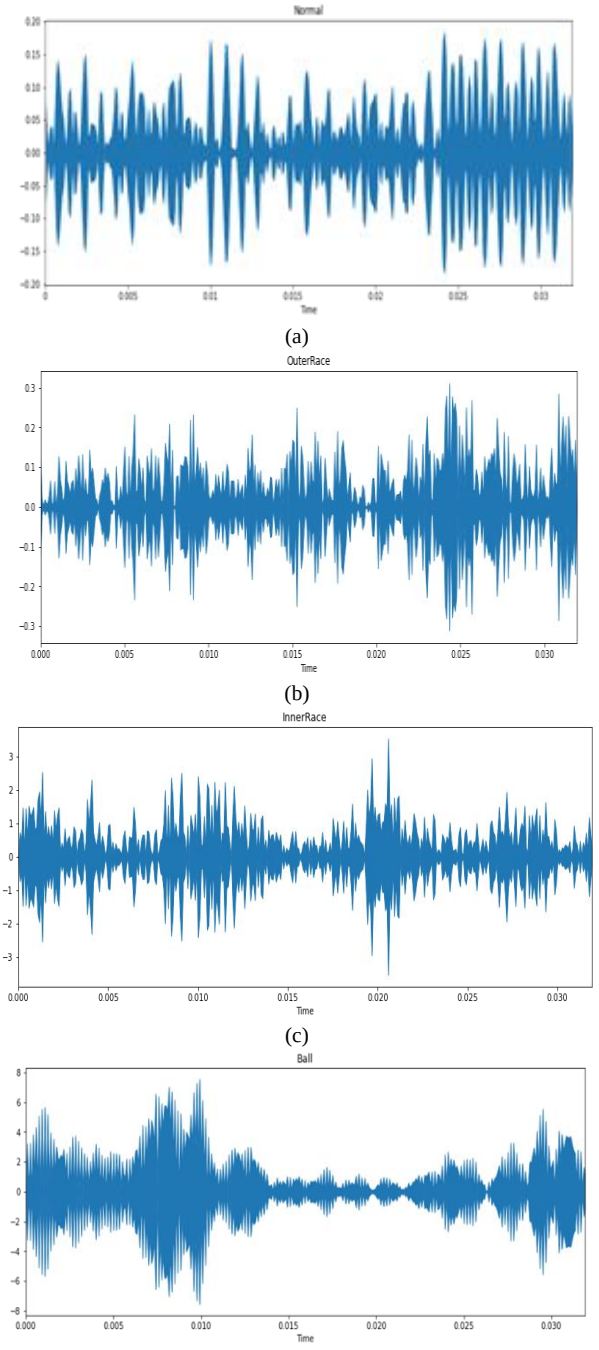


Fig. 6. (a). Time-domain Feature Representation for Normal Bearing, (b). Time-Domain Feature Representation for Outer Race Fault, (c). Time-Domain Feature Representation for Inner Race Fault, (d). Time-Domain Feature Representation for Ball Fault.

F. Parameter Settings for CNN-Bi-LSTM Model

The summary of the model's parameters set for the proposed hybrid CNN-Bi-LSTM architecture is shown in Fig. 7.

The first four convolutional layers used batch normalization and spatial dropout with a value of 0.25, which could effectively improve the performance of the network by preventing the overfitting problem. Softsign activation function was used as a classifier in Bi-LSTM with filter size of 256 and Adam optimizer for compilation. The loss function categorical cross-entropy and batch size of 32 was set to identify the fault state by setting a short time of 100 epochs.

The trainable parameters are those which are learnt by the model during the feature learning from the classification layers namely convolution, LSTM, and the fully connected layers. The non-trainable parameters are learnt by the model from the batch normalization layers.

Layer (type)	Output Shape	Param #
conv1d_4 (Conv1D)	(None, 382, 128)	512
leaky_re_lu_7 (LeakyReLU)	(None, 382, 128)	0
max_pooling1d_4 (MaxPooling 1D)	(None, 191, 128)	0
batch_normalization_7 (Batc hNormalization)	(None, 191, 128)	512
spatial_dropout1d_4 (Spatia lDropout1D)	(None, 191, 128)	0
conv1d_5 (Conv1D)	(None, 189, 256)	98560
leaky_re_lu_8 (LeakyReLU)	(None, 189, 256)	0
max_pooling1d_5 (MaxPooling 1D)	(None, 94, 256)	0
batch_normalization_8 (Batc hNormalization)	(None, 94, 256)	1024
spatial_dropout1d_5 (Spatia lDropout1D)	(None, 94, 256)	0
conv1d_6 (Conv1D)	(None, 92, 512)	393728
leaky_re_lu_9 (LeakyReLU)	(None, 92, 512)	0
max_pooling1d_6 (MaxPooling 1D)	(None, 46, 512)	0
batch_normalization_9 (Batc hNormalization)	(None, 46, 512)	2048
spatial_dropout1d_6 (Spatia lDropout1D)	(None, 46, 512)	0
conv1d_7 (Conv1D)	(None, 45, 256)	262400
leaky_re_lu_10 (LeakyReLU)	(None, 45, 256)	0
max_pooling1d_7 (MaxPooling 1D)	(None, 45, 256)	0
batch_normalization_10 (Bat chNormalization)	(None, 45, 256)	1024
spatial_dropout1d_7 (Spatia lDropout1D)	(None, 45, 256)	0
bidirectional_1 (Bidirectio nal)	(None, 45, 512)	1050624
leaky_re_lu_11 (LeakyReLU)	(None, 45, 512)	0
batch_normalization_11 (Bat chNormalization)	(None, 45, 512)	2048
dropout_3 (Dropout)	(None, 45, 512)	0
lstm_3 (LSTM)	(None, 256)	787456
leaky_re_lu_12 (LeakyReLU)	(None, 256)	0
batch_normalization_12 (Bat chNormalization)	(None, 256)	1024
dropout_4 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 64)	16448
leaky_re_lu_13 (LeakyReLU)	(None, 64)	0
batch_normalization_13 (Bat chNormalization)	(None, 64)	256
dropout_5 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 12)	780

Total params: 2,618,444
Trainable params: 2,614,476
Non-trainable params: 3,968

Fig. 7. Summary of Model's Parameters.

G. Confusion Matrix

The confusion matrix is the representation of matching degree between the actual and predicted labels in the form of matrix. The confusion matrix for the proposed 1DCNN-Bi-LSTM model is shown in Fig. 8. The model has correctly classified 4111 samples out of 4118 by demonstrating an accuracy of 99.84%.

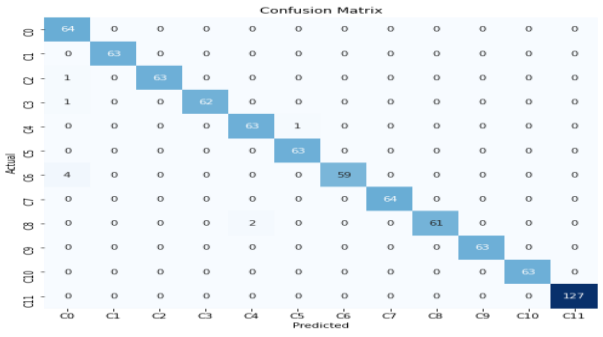


Fig. 8. Confusion Matrix.

H. Learning Curve

A learning curve is a plot of model's learning performance over experience or time. The model is evaluated during training phase after each update based on the training and validation dataset. It gives an idea of how well the model is learning and generalizing. The learning curve for the proposed method is shown in Fig. 9. It demonstrates the training and validation accuracy versus number of epochs.

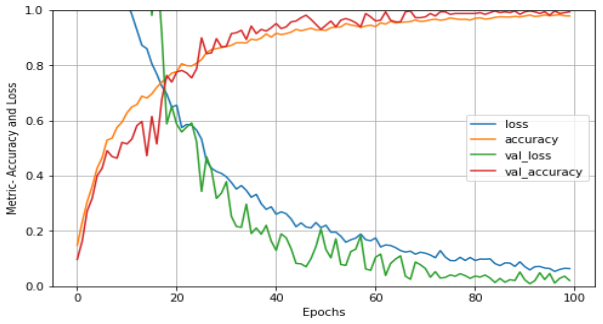


Fig. 9. Learning Curve for Hybrid CNN-Bi-LSTM Model.

A comparative analysis of the proposed hybrid model is made with other existing DL models as shown in Table II. The performance indicates that the proposed model accomplishes better results in classifying multiple faults as compared to other models.

TABLE II. COMPARATIVE ANALYSIS OF THE PROPOSED METHOD WITH OTHER DL MODELS

Model	Training Accuracy	Test Accuracy	Validation Accuracy
1D-CNN only	98.18	95.99	95.45
1D-CNN with LSTM	98.18	99.51	99.09
1D-CNN-Bi-LSTM	98.01	98.78	99.09
1D-CNN-Bi-LSTM (Softsign and Spatial Dropout)	99.84	98.17	99.69

VI. CONCLUSION

In this research, A Hybrid 1D-CNN-Bi-LSTM model with Spatial Dropout for Multiple Fault Diagnosis of Roller Bearing is proposed. Usage of Spatial Dropout technique and Softsign activation function in the proposed hybrid fault diagnosis method has shown an improvement in the accuracy by performing automatic feature extraction and preventing the problem of overfitting. 1D-CNN extracts the features from the raw signal and Bi-LSTM layer fuses the feature information to enhance stability of the model and classify the faults. The efficiency of the proposed model is analysed by considering CWRU bearing vibration data for experimentation. A comparative analysis of the proposed method is made with other existing models. The model has shown the performance accuracy of 99.84% in classifying 12 different fault types. Therefore, the proposed hybrid 1D-CNN-Bi-LSTM (Softsign and Spatial Dropout) is an effective multi-class fault diagnosis method with the prevention of model overfitting problem.

REFERENCES

[1] Yuan, Laohu, Dongshan Lian, Xue Kang, Yuanqiang Chen, and Kejia Zhai. "Rolling bearing fault diagnosis based on convolutional neural network and support vector machine." IEEE Access 8 (2020): 137395-137406.

[2] Gupta, Pankaj, and M. K. Pradhan. "Fault detection analysis in rolling element bearing: A review." Materials Today: Proceedings 4, no. 2 (2017): 2085-2094.

[3] Neupane, Dhiraj, and Jongwon Seok. "Bearing fault detection and diagnosis using case western reserve university dataset with deep learning approaches: A review." IEEE Access 8 (2020): 93155-93178.

[4] Zhang, Shen, Shibo Zhang, Bingnan Wang, and Thomas G. Habetler. "Deep learning algorithms for bearing fault diagnostics—A comprehensive review." IEEE Access 8 (2020): 29857-29881.

[5] Tang, Shengnan, Shouqi Yuan, and Yong Zhu. "Deep learning-based intelligent fault diagnosis methods toward rotating machinery." Ieee Access 8 (2019): 9335-9346.

[6] Chen, Zhiqiang, Shengcai Deng, Xudong Chen, Chuan Li, René-Vinicio Sanchez, and Huafeng Qin. "Deep neural networks-based rolling bearing fault diagnosis." Microelectronics Reliability 75 (2017): 327-333.

[7] Lei, Jinhao, Chao Liu, and Dongxiang Jiang. "Fault diagnosis of wind turbine based on Long Short-term memory networks." Renewable energy 133 (2019): 422-432.

[8] Chen, Zhuyun, and Weihua Li. "Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network." IEEE Transactions on Instrumentation and Measurement 66, no. 7 (2017): 1693-1702.

[9] Zhang, Yuyan, Xinyu Li, Liang Gao, Wen Chen, and Peigen Li. "Intelligent fault diagnosis of rotating machinery using a new ensemble deep auto-encoder method." Measurement 151 (2020): 107232.

[10] Zhao, Rui, Dongzhe Wang, Ruqiang Yan, Kezhi Mao, Fei Shen, and Jinjiang Wang. "Machine health monitoring using local feature-based gated recurrent unit networks." IEEE Transactions on Industrial Electronics 65, no. 2 (2017): 1539-1548.

[11] Ince, Turker, Serkan Kiranyaz, Levent Eren, Murat Askar, and Moncef Gabbouj. "Real-time motor fault detection by 1-D convolutional neural networks." IEEE Transactions on Industrial Electronics 63, no. 11 (2016): 7067-7075.

[12] Abed, Wathiq, Sanjay Sharma, Robert Sutton, and Amit Motwani. "A robust bearing fault detection and diagnosis technique for brushless DC motors under non-stationary operating conditions." Journal of Control, Automation and Electrical Systems 26, no. 3 (2015): 241-254.

[13] Zou, Ping, Baocun Hou, Jiang Lei, and Zhenji Zhang. "Bearing fault diagnosis method based on EEMD and LSTM." International Journal of Computers Communications & Control 15, no. 1 (2020).

[14] Cao, Lixiao, Zheng Qian, Hamidreza Zareipour, Zhenkai Huang, and Fanghong Zhang. "Fault diagnosis of wind turbine gearbox based on

deep bi-directional long short-term memory under time-varying non-stationary operating conditions." IEEE Access 7 (2019): 155219-155228.

[15] Choudakkanavar, Gangavva, J. Alamelu Mangai, and Mohit Bansal. "MFCC based ensemble learning method for multiple fault diagnosis of roller bearing." International Journal of Information Technology (2022).

[16] Wang, Zheng, Hongzhan Ma, Hansi Chen, Bo Yan, and Xuening Chu. "Performance degradation assessment of rolling bearing based on convolutional neural network and deep long-short term memory network." International Journal of Production Research 58, no. 13 (2020): 3931-3943.

[17] Sun, Haibin, and Shichao Zhao. "Fault diagnosis for bearing based on 1DCNN and LSTM." Shock and Vibration 2021 (2021).

[18] Zhang, Xiaochen, Yiwen Cong, Zhe Yuan, Tian Zhang, and Xiaotian Bai. "Early fault detection method of rolling bearing based on MCNN and GRU network with an attention mechanism." Shock and Vibration 2021 (2021).

[19] Jin, Guoqiang, Tianyi Zhu, Muhammad Waqar Akram, Yi Jin, and Changan Zhu. "An adaptive anti-noise neural network for bearing fault diagnosis under noise and varying load conditions." IEEE Access 8 (2020): 74793-74807.

[20] Yu, Wenbing, and Pin Lv. "An end-to-end intelligent fault diagnosis application for rolling bearing based on MobileNet." IEEE Access 9 (2021): 41925-41933.

[21] Gu, Kai, Yu Zhang, Xiaobo Liu, Heng Li, and Mifeng Ren. "DWT-LSTM-based fault diagnosis of rolling bearings with multi-sensors." Electronics 10, no. 17 (2021): 2076.

[22] Zhi, Zhuo, Liansheng Liu, Datong Liu, and Cong Hu. "Fault detection of the harmonic reducer based on CNN-LSTM with a novel denoising algorithm." IEEE Sensors Journal 22, no. 3 (2021): 2572-2581.

[23] Huang, Ting, Qiang Zhang, Xiaoan Tang, Shuangyao Zhao, and Xiaonong Lu. "A novel fault diagnosis method based on CNN and LSTM and its application in fault diagnosis for complex systems." Artificial Intelligence Review 55, no. 2 (2022): 1289-1315.

[24] Hotait, Hassane, Xavier Chiementin, and Lanto Rasolofondraibe. "Intelligent Online Monitoring of Rolling Bearing: Diagnosis and Prognosis." Entropy 23, no. 7 (2021): 791.

[25] Luo, Jiahang, and Xu Zhang. "Convolutional neural network based on attention mechanism and Bi-LSTM for bearing remaining life prediction." Applied Intelligence 52, no. 1 (2022): 1076-1091.

[26] https://engineering.case.edu/bearingdatacenter

[27] Li, Mingyong, Qingmin Wei, Hongya Wang, and Xuekang Zhang. "Research on fault diagnosis of time-domain vibration signal based on convolutional neural networks." Systems Science & Control Engineering 7, no. 3 (2019): 73-81.

[28] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In International conference on machine learning, pp. 448-456. PMLR, 2015.

[29] Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15, no. 1 (2014): 1929-1958.

[30] Pan, Honghu, Xingxi He, Sai Tang, and Fanming Meng. "An improved bearing fault diagnosis method using one-dimensional CNN and LSTM." Strojniski Vestnik/Journal of Mechanical Engineering 64 (2018).

[31] Qiu, Dawei, Zichen Liu, Yiqing Zhou, and Jinglin Shi. "Modified Bi-directional LSTM neural networks for rolling bearing fault diagnosis." In ICC 2019-2019 IEEE international conference on communications (ICC), pp. 1-6. IEEE, 2019.

[32] https://www.tensorflow.org/api_docs/python/tf/keras

AUTHORS' AFFILIATIONS

First Author received her B.E. and M.Tech degree in Computer Science and Engineering from Visvesvaraya Technological University (VTU), Belagavi, in 2009 and 2013 respectively. Currently pursuing Ph.D. degree from Presidency University, Bangalore. Since from 2013, she is working as Assistant Professor in various recognized institutions. Her research interests include Machine Learning, Deep Learning and Fault Diagnosis.

Second Author received her Ph.D from BITS Pilani, Dubai Campus in 2015. She is working in the department of Computer Science and Engineering at Presidency University, Bangalore. Her research interests include Data Mining and Machine Learning Algorithms and Applications.

Building and Testing Fine-Grained Dataset of COVID-19 Tweets for Worry Prediction

Tahani Soud Alharbi, Fethi Fkih
Department of Computer Science, College of Computer
Qassim University, Buraydah 51452, Saudi Arabia

***Abstract*—The COVID-19 outbreak has resulted in the loss of human life worldwide and has increased worry concerning life, public health, the economy, and the future. With lockdown and social distancing measures in place, people turned to social media such as Twitter to share their feelings and concerns about the pandemic. Several studies have focused on analyzing Twitter users' sentiments and emotions. However, little work has focused on worry detection at a fine-grained level due to the lack of adequate datasets. Worry emotion is associated with notions such as anxiety, fear, and nervousness. In this study, we built a dataset for worry emotion classification called "WorryCov". It is a relatively large dataset derived from Twitter concerning worry about COVID-19. The data were annotated into three levels ("no-worry", "worry", and "high-worry"). Using the annotated dataset, we investigated the performance of different machine learning algorithms (ML), including multinomial Naïve Bayes (MNB), support vector machine (SVM), logistic regression (LR), and random forests (RF). The results show that LR was the optimal approach, with an accuracy of 75%. Furthermore, the results indicate that the proposed model could be used by psychologists and researchers to predict Twitter users' worry levels during COVID-19 or similar crises.**

***Keywords*—COVID-19; sentiment analysis; emotion analysis; worry dataset; concern analysis**

I. Introduction

At the end of the year 2019, China reported cases of pneumonia caused by an unknown virus in Wuhan City. Later, this pneumonia was defined by the World Health Organization (WHO) as the coronavirus disease 2019 (COVID-19)[1]. It was then declared a pandemic that has had multiple consequences, including the death and long-term effects of infected people. According to WHO, as of July 2022, the total number of reported COVID-19 cases was approximately 545 million, with a total of 6.3 million deaths[1]. The uncertainty and low predictability of COVID-19 threaten people's both physical and mental health, especially in terms of emotions and cognition [2]. The most challenging effects of the pandemic, especially during lockdowns, are depression, anxiety, and worries due to unemployment, losing loved ones, or being personally affected by the disease [3]. While there are several programs that psychologists and therapists carry out to enable recovery from these issues, there is an immense need to study worry using other sources [4]. Traditional methods of public health monitoring, like questionnaires and clinical tests, have certain limitations; for example, they only cover a limited number of participants and are restricted to the data collection period[5].

In contrast, social media are becoming a significant source of rich real-time information during crises, including disease outbreaks and natural disasters [6]. Twitter is a unique source of big data for public health researchers due to the real-time nature of the content and the ease of searching and accessing publicly available data [7]. In this vein, COVID-19-related behaviors and sentiments are available on social media. Twitter users continuously post about their feelings and worries regarding these unusual circumstances[8]. This situation drew the attention of computer scientists and researchers, leading to numerous studies on the understanding of the emotional states during current events, especially those related to the pandemic [9].

The research problem is related to the discrimination of the worry analysis studies. Most of the researchers have focused on discrete emotion theories, like Ekman's emotion classification schema [10], by annotating texts to the six basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise) [11]. As the most dominant emotions during crises are worry and anxiety [12], [13], the existing methods for emotion detection are insufficient to capture the emotion of worry accurately [14].

Detecting worry is complex as people are either unwilling to disclose worries to medical personnel or prefer sharing their feelings on social media. Thus, there is a lack of datasets that could be used for worry analysis, as many studies depend on surveys and interviews [15], [16]. To the best of the authors' knowledge, this is the first study to build a to-date dataset about COVID-19-related worries that is to be applied to machine learning (ML) models. In the context of this paper, worry about COVID-19 is classified into three fine-grained levels: "no-worry", "worry", and "high-worry". The "no-worry" category includes people discussing the news and politics about the virus or content-containing statistics and figures. On the other hand, people expressing high levels of feelings such as panic or fear ("high worry" category) are distraught. Between these two categories ("worry" category), there are people expressing concern about the virus, who are considered stressed about the present and the future.

The contribution of this paper is two-fold. First, the WorryCov[2] dataset was built based on three classes: "no-worry", "worry", and "high-worry". It was built with experts

[1]https://covid19.who.int

[2]Dataset is available from the authors upon reasonable request.

in linguistics and followed an annotation scheme under strict quality control. Then, several ML classification models were used to test the dataset.

The paper is outlined as follows. The related works are discussed in Section 2. Section 3 introduces the proposed approach. Section 4 provides the results and discussion, while Section 5 concludes the paper.

II. Related Work

Worry analysis is considered one dimension of emotion analysis frequently studied in the literature. Therefore, this study focuses on concern, sentiment, and emotion analysis towards or during disasters or pandemics such as the COVID-19 pandemic.

Much previous research was carried out to determine the public health concerns toward disasters or epidemics based on sentiment analysis results. For example, the work in [17] aimed to analyze Twitter messages relating to Hurricane Irene and trained a dataset based on sentiment analysis classifiers to categorize tweets into levels of concern. They evaluated the impact of various tokenization strategies and feature choices like a bag of words (BOW) and lexicons on classification accuracy. With 84.27% accuracy, the best settings for the maximum entropy classifier were removing punctuation, converting the text to lowercase, removing stop words, and building a worry lexicon. The Epidemic Sentiment Monitoring System [5] provides visualization tools for Twitter posts responding to public concerns about different diseases. The degree of concern reported that multinomial Naïve Bayes (MNB) achieved the highest F1-score using term frequency-inverse document frequency (TF-IDF) features. To measure and monitor public health concerns about communicable diseases, a sentiment classification approach was applied to Twitter data by measuring different levels of concern [18]. The classifier was trained with a dataset automatically generated by a programming system using an emotion-oriented and clue-based method. Three ML classifiers were evaluated, with the NB classifier achieving the best accuracy for the epidemic-related dataset.

Regression is often used to detect public health concerns. For instance, in [19], a strategy to predict to what extent news about a public health issue can be disseminated was proposed using a data collection of microblog news posts. This ML method relies on the logistic regression (LR) algorithm that automatically categorized news posts into two classes: normal news or news posts that resonated with widespread public anxiety.

As for COVID-19, abundant works have already been published studying the effects of this pandemic on various aspects. For example, most research focused on analyzing Twitter data and finding the main critical topics that raise concerns for individuals regarding the COVID-19 pandemic. In [20], [21] used the topic modeling technique LDA (an unsupervised machine learning model) to identify the most common topics in the tweets and performed sentiment analysis. Furthermore, analyzing citizens' concerns during the COVID-19 epidemic has been studied in [22]. 30,000 COVID-19-related tweets were collected from March 14, 2020. Each tweet was labeled as very negative, negative, neutral, very positive, and positive by using the natural language processing (NLP) library. Then, the authors used sentiment analysis on pre-processed tweets to show the level of concern in various US states. They presented an approach for measuring citizens' concern levels through Twitter data by using the ratio of very negative and negative tweet counts over the total number of tweets in the dataset. As a result, school closing-related tweets cause the highest level of concern among citizens. Similarly, the study [23] presented a method to identify the COVID-19 topic's degree of concern through user conversations on Twitter based on two phases of the classification process. The first classification step is to separate tweets into two classes, namely COVID-19 and non-COVID-19. The second step is to classify the COVID-19 data into seven topics: donations, emotional support, warnings and suggestions, hoaxes, notification of information, seeking help, and criticism. Six pairs of combinations of word-level and character-level word embeddings, namely Word2Vec and fastText, with three deep learning models, CNN, RNN, and LSTM, were used to apply the text classification model. The best accuracy was achieved when fastText and LSTM were used together for both stages of classification, with 97.3% and 99.4%, respectively.

Significant research in public health has applied emotion analysis using social media-derived information to monitor public emotions during disease outbreaks. Emotions such as anxiety, anger, happiness, desire, disgust, fear, relaxation, and sadness have been widely studied. Emotions are often linked with topic modeling to identify the topics and their intensity level. For example, findings in [12] indicate that the longer texts gave insights into what people worry about during the pandemic: the economy and the family. In the SenWave system [8], seven fine-grained sentiment categories, namely, optimistic, thankful, empathetic, pessimistic, anxious, sad, annoyed, denial, official, and joking, are used to study the concern of Twitter users from different countries. The labeled tweets are used to train the deep learning language models such as XLNet, AraBert, and ERNIE, while over 105 million unlabeled tweets are used for the testing process. An XLNet pre-trained language model was used for English tweets. The classifier achieved an 80% accuracy, which proves the efficiency of the models. However, emotion analysis studies are minimal compared with sentiment research due to the lack of annotated data [24]. The EmoBERT model [24] was used to capture emotions related to emotional health (annoyed, anxious, empathetic, sad) to compare emotions expressed on social media before and during the COVID-19 epidemic. In comparison to BERT and XLNet, EmoBERT achieved better results.

Our review shows that little research has addressed worry detection. However, many studies address anxiety as an issue of mental health, for instance, this study [25]utilized personal narratives from Reddit to detect anxiety disorders and classified anxiety-related posts into a binary level of anxiety. Using various linguistic features, including vector-space representations (Word2Vec and Doc2Vec), topic (LDA) models, Linguistic Inquiry and Word Count (LIWC) dictionary, and n-gram language models. Overall, all features

that have been used succeeded in classifying the level of anxiety, for single-source features, using Neural Network with N-gram probabilities achieved slightly better accuracy (92%) compare with using SVM with word-vector embeddings (word2vec), and for combined features, Neural Network has produced the highest accuracy of 98% by aggregating LIWC with word2vec embeddings and by aggregating N-gram features with LIWC. Moreover, this paper [26] developed its own binary classification dataset for detecting anxiety and depression users on social media who have not yet been diagnosed with mental illness. The authors have presented a comparative experimental evaluation using the traditional linear model and pre-trained LMs (language models). Their results showed that LMs (BERT and ALBERT) performed relatively well with balanced training data. However, in unbalanced training sets, Support Vector Machine (SVM) with word embeddings and TF-IDF features performed slightly better overall, with 0.750 F1-score, 0.747 for accuracy, and 0.740 for precision.

To our knowledge, Verma et al.'s study [14] is the most relevant to the prediction of worry using Twitter data. Using crowdsourcing, they re-annotated an existing dataset that contains four emotions (joy, anger, fear, and sadness) [27] for worry classification. A wide range of machine learning and deep learning models were evaluated. For traditional ML approaches, Multinomial Naive Bayes (MNB), and Support Vector Machine (SVM) are implemented by using feature-based models. For deep learning based on word embeddings, they used Hierarchical Attention Network (HAN) and CNN-static with combined Glove emoji2vec embeddings. While deep learning approaches based on contextual embeddings were also applied like RoBERT and XLNet. The results showed that deep learning methods outperform as compared to the traditional models for worry identification with 0.61 F1-score.

The gap in the current studies is related to the lack of a large new dataset for worry level detection related to COVID-19 tweets. Despite the many works, most previous results focus on the sentiment classification of tweets as positive, negative, and natural.

III. Proposed Approach

The proposed approach is shown in Fig. 1. Due to the limited dataset related to worry identification from text, we built a dataset and chose the classification task. We decided to select some machine learning models to validate the credibility of the collected dataset. The approach first described the data collection and the annotation process into three levels of worry using COVID-19-related tweets. The dataset was then used to extract features, run ML models, and evaluate the results.

A. Building the Benchmark Dataset

1) Dataset collection and filtering: To build the benchmark dataset, tweets were collected, filtered, and annotated. Twitter is one of the most popular social media and has a wide range of content including rich text, emojis, and hashtags [14]. The tweets related to COVID-19 were collected using Tweepy, the Python Twitter API library [28]. Initially, we used unified query keywords (i.e., coronavirus, covid-19, #coronavirus, and #covid-19), previously used in other studies [29], to identify the tweets related to COVID-19. The tweets were collected over three periods to ensure that they covered significant milestones during the pandemic. The three periods are consistent with [30] and are the following:

- First period: from January 30 to February 28, 2020. During this period, the first COVID-19-induced death was reported in China, and WHO announced a public health emergency.
- Second period: from March 29 to April 29, 2020. During this period, WHO declared COVID-19 a worldwide pandemic, leading many governments to impose restrictions on citizens in an attempt to reduce the spread of the virus.
- Third period: from May 10 to June 30, 2020. During this period, COVID-19 had spread globally, with an increased number of confirmed cases and deaths.

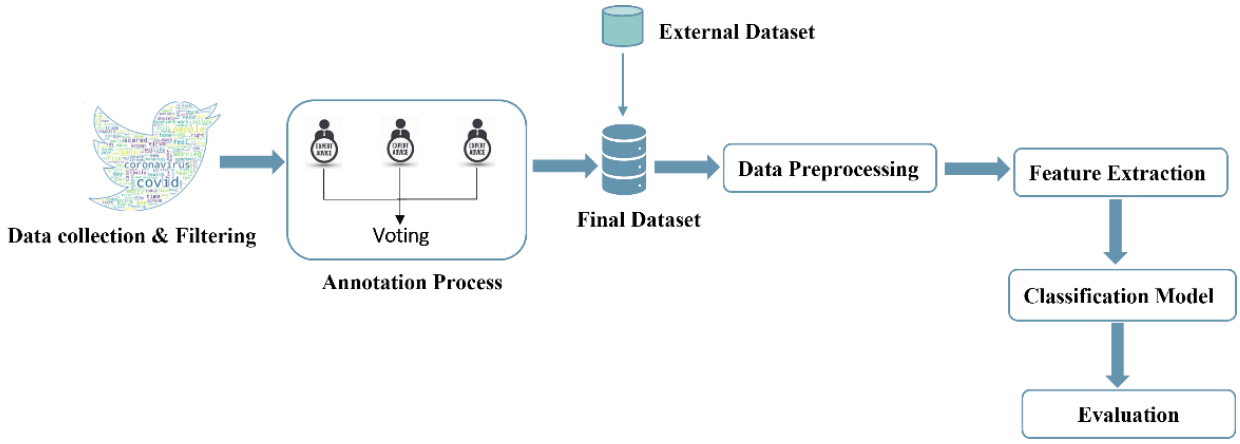


Fig. 1. Proposed Approach.

Following these periods and using the aforementioned keywords, 270,000 tweets were collected. Each tweet had 24 columns, including data and time, username, tweet text, and location. Since we wanted to detect feelings of worry at the tweet level, we removed the rest of the columns and only retained the text column. However, a large proportion of COVID-19-related tweets were probably not associated with one emotion; thus, annotating them would be costly and ineffective [31], [32]. To meet our objective, we focused solely on the worry emotion and used worry-related keywords to create a dataset of tweets representing this emotion. Following [33], we selected keywords (terms) to filter the collected data. The terms were extracted from Thesaurus.com by finding synonyms and terms related to worry; the dictionary is one of the trusted, free online dictionaries. The synonym keywords are shown in Table I.

Often, datasets contain noise and irrelevant text. Therefore, the following rules were applied to reduce the dataset to more concise and related tweets: (1) deleting duplicate tweets (i.e., retweeted by other users), (2) deleting non-English language tweets, and (3) deleting all tweets less than 40 characters (short tweet).

TABLE I. KEYWORDS USED FOR FILTERING TWEETS (EXTRACTED FROM THE ONLINE THESAURUS.COM DICTIONARY)

worry	anxiety	concern	apprehension	fear	afraid
worried	anxious	panic	stress	tension	terrify
worries	distress	nervous	uncertain	tense	terrified
scary	confusion	restless	doubt	horrible	terror
scared	confuse	pressure	uptight	horror	paranoid
discomfort	troubled	pain	upset	dread	alarm

2) Annotation process: Manual labeling of social media data is challenging and requires dedicated time from domain experts (time-consuming). However, it is a critical part of the data preparation process in supervised learning. We annotated the data for not just coarse classes (such as worry or no-worry) but also for fine-grained levels indicating the intensity or degrees of emotion. However, annotating instances for degrees of emotions is a more difficult task to ensure annotation consistency [33]. Therefore, this study followed a set of rules to overcome this challenge: (1) tweets were annotated to three classes only: "no-worry", "worry", and "high-worry",(2) Three English speakers with more than three years of experience in linguistics were employed; (3) the majority vote was used to annotate an individual tweet, and when the three experts disagreed, the tweet was considered irrelevant and was removed from the dataset; moreover, a newly developed website application was used to help the annotators accomplish their work; and (4) each annotator got the same number of tweets (2,700) for each month of the three periods (8,100 tweets in total). This process was slow but ensured results in accordance with the following guidelines to classify each tweet:

- "No-worry" class:
  - News or politics (i.e., conspiration theories, China-related discussion where the Chinese are blamed for the virus, US politics, critics of Donald Trump, etc.) and facts (e.g., numbers, statistics).

    Example: "China's outbreak is serious. But flu killed *5000 Americans* in the 1st 2wks of 2020 coronavirus infected 6, killed 0. Not sure what info you have that CDC director doesn't "The immediate risk to the US public is low." Our US readers deserve to know they don't need to panic."
  - Other diseases (i.e., tweets comparing COVID with other diseases, discussing symptoms and mortality rates).

    Example: "Aids is a killer disease Cancer is a killer disease Ebola is a killer disease Swine flu is a killer disease The only thing that divides Coronavirus to this other diseases is the fact that it is just the latest, stop with the panic and take care of yourself! #coronavirus."
  - Expressing some other emotions (i.e., tweets denying the existence of the virus or expressing any optimistic/positive attitude toward it).

    Example: "markets are full of pads soap Dettol, etc. People are not freaking out as they know there's enough. They aren't crazy buying. Let's hope the panic ends soon all over the world and we live happily again #covid-19."
- "Worry" class:
  - Expressing general concern (i.e., mentioning being worried/stressed/concerned about the present and future of COVID-19).

    Example: "Also, I'm young and healthy and unlikely to die from covid-19, so no reason to be afraid at all for me. I'm nervous about infecting those who are less likely to survive though, so I will do my best to prevent that of at all possible if I get infected."
- "High-worry" class:
  - Expressing concern (i.e. tweets expressing feelings of panic, fear, etc.).

    Example: "i'm tired of crying. i'm tired of the anxiety, and panic attacks. i want to go outside again. please - STAY HOME. #COVID-19 #COVID19Ontario."
  - Frequent use of intensifiers (e.g., extremely, so, very) and featuring content related to (fear of) death.

    Example: "So much stress, so much anxiety, AND I'M PREGNANT. Headaches all day, puking many times a day, quarantined. People are dying, this is not cool. #coronavirus"

The annotation resulted in 7,861 instances corresponding to the three classes. The "no-worry" class included 3,158 instances, the "worry" class had 3,127 instances, and the "high-worry" class included 1,576 instances. The remaining 239 tweets were eliminated as the annotators disagreed with their classification (not sure). However, we noticed that the WorryCov dataset is imbalanced. So, it should be solved to reduce skewness and increase the performance of ML models [34]. Therefore, we decided to expand the dataset using other external datasets. To our knowledge, no dataset focuses on only worry emotion. Therefore, we selected the intensity of anxiety based on [11] since it was considered a synonym for worry. Anxiety levels in [11] ranged from 1 to 9, where 1 was considered the lowest and 9 the highest. Considering this range, we chose the intensity levels 7, 8, and 9 as descriptive of the "high-worry" class, resulting in a total of 3,127 instances in the "high-worry" class.

B. Prediction of Worry Levels

The balanced benchmark dataset was used to evaluate the performance of different ML models. In this section, the data preprocessing, feature extraction, classification, and evaluation steps of this dataset are discussed.

1) Data preprocessing: Preprocessing generally improves the data quality by extracting meaningful fragments from a given text excluding the noise [35], [36]. Preprocessing steps include text cleaning such as URL, digit, punctuation removal, etc., and lemmatization.

In the cleaning step, we removed URLs, user mentions, and hashtags. Previous research on sample datasets shows that these items do not provide any evidence of the level of worry in tweets or useful information [37]. Next, each tweet was converted to lowercase to avoid considering the exact words as unique features, such as "HELP", "Help", or "help" will be converted to "help" [38]. Then, the contractions (i.e., "I'm" instead of "I am") were replaced by the original phrase as described in [37]. Next, digits, punctuation marks, and extra spaces that do not provide any semantic information to the text were removed. NLP classification tasks often involve removing stop words to improve performance metrics [39]. However, in this dataset, worry feelings were frequently expressed as ideas about oneself, leading to the use of the "I" and "my" pronouns. Therefore, stop words were not removed to retain the linguistic characteristics of worried users. Finally, each word was lemmatized using Wordnet Lemmatizer available in the natural language toolkit (NLTK) library [40].

2) Feature extraction: Often called a features vector [37], this step refers to transforming raw data into numerical data that machines can understand. Term frequency-inverse document frequency (TF-IDF) is a popular text vectorization technique to generate vector representations of a text [41] and was employed in this experiment. The TF-IDF weighting scheme is based on two parts: term frequency (TF) and inverse document frequency (IDF). TF-IDF is mathematically formulated in the following (1) [42]:

$$\text{TF} - \text{IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \quad (1)$$

where t denotes a term and d denotes a document.

TF is the frequency of any term within a given document and is calculated by dividing the number of mentions of a given word by the total number of words in the document [37], TF is defined by (2) [43]:

$$\text{TF}(t, d) = \text{Number of times the term } t \text{ appears in the document / Total number of terms in the document} \quad (2)$$

IDF represents the importance of a term in the corpus of the text. It is a technique that combined with TF reduces the impact of common words. There are some words, like "the", "is", "and", etc., that occur frequently but are void of information. IDF is defined by Eq. (3) [44]:

$$\text{IDF}(t) = \log(\text{Number of documents / Documents containing the term t}) \quad (3)$$

C. ML-Based Classifiers

Four ML-based classifiers were used in the multi-classification task. These methods were multinomial NB (MNB), logistic regression (LR), Support Vector Machine (SVM), and Random Forest (RF). The default settings of these methods were taken from the scikit-learn library [45]. MNB is suitable for classifying discrete features or fractional counts such as TFIDF. LR calculates the likelihood of a target variable based on a collection of independent variables and a given dataset. SVM is a classification algorithm for two-group classification problems (in our case one-vs-rest scheme is used). Finally, the RF algorithm builds many random decision trees using bagging and feature randomness for each tree.

D. Evaluation Metrics

Each classifier was evaluated using the following performance measurements: accuracy, precision, recall, and F1-score. These standard metrics are defined as follows:

Accuracy is the ratio of the number of correct predictions to the overall number of predictions:

$$\text{Accuracy} = (TN + TP) / (TN + TP + FP + FN) \quad (4)$$

Precision is the ratio of the correctly predicted positive instances to the total positive instances:

$$\text{Precision (P)} = TP / (TP + FP) \quad (5)$$

Recall is the ratio of the correctly predicted positive instances to the total of all instances in the actual class:

$$\text{Recall (R)} = TP / (TP + FN) \quad (6)$$

F1-score is the harmonic average of precision and recall:

$$\text{F1-score} = (2 \times P \times R) / (P + R) \quad (7)$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

IV. Results and Discussion

After filtering (see Section 3.1), we obtain 15,000 tweets. Fig. 2 presents the word cloud of the most commonly used words in the WorryCov dataset. The most frequent keywords are related to the COVID-19 pandemic, such as "Covid",

"corona", "coronavirus", and "scared". The distribution of tweets among the three worry levels is shown in Fig. 3. The figure demonstrates that the three classes are balanced. Fig. 4 includes tweets representative of the three levels of worry. The figure shows that the "high-worry" class shows fear and stress behavior. While the "worry" class indicates familiar people's behavior during any pandemic. In contrast, the "no-worry" indicates informative or news content or an optimistic feeling.

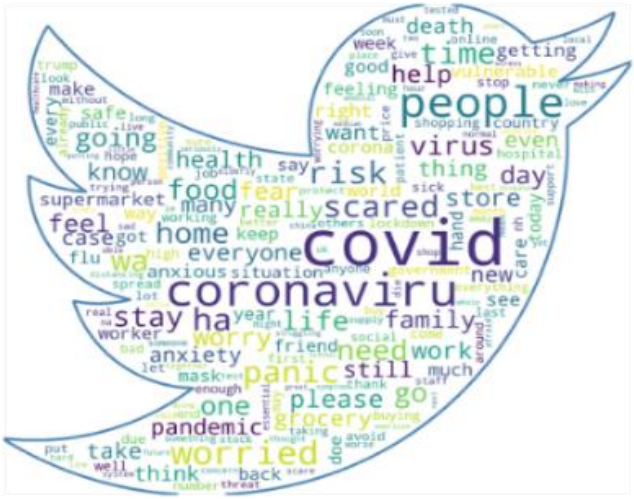


Fig. 2. The Word Cloud of the Collected Tweets.

Fig. 3. Distribution of the Tweets among the Three Worry Levels.

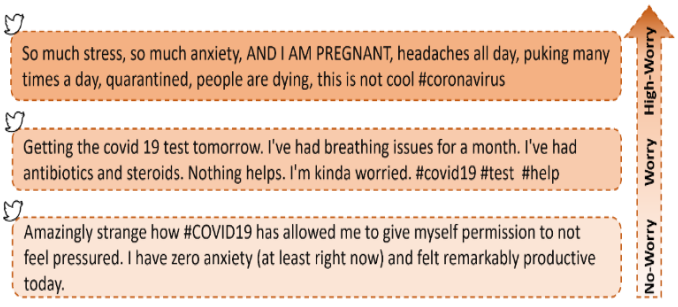


Fig. 4. A Sample of Tweets from the Benchmark Dataset Representing the Three Levels of Worry.

To predict the performance of the selected ML models, the dataset was split into 80% for training and 20% for testing. Next, data preprocessing and feature extraction (see Section 3.2) were employed to extract relevant features. Finally, the accuracy, precision, recall, and F1-score results were reported for the average class (Table II).

As shown in Table II, the classification performance of LR (reported in bold) performed better than the other models in terms of accuracy, precision, recall, and F1-score. It yielded the highest accuracy of 75%, a precision of 0.751, a recall of 0.747, and an F1-score of 0.748. On the contrary, RF acquired the lowest values with an accuracy of 68%, 0.683 for recall, 0.682 for precision, and F1-score of 0.682.

An investigation of the dataset shows that the LR algorithm was able to build features better than others because the training algorithm of LR uses the one-vs-rest scheme in the multiclass option and the cross-entropy loss. However, the results cannot be generalized as the absolute difference among the high-performing models in Table II is less than 4%.

Fig. 5, 6, and 7 show the precision, recall, and F1-score measurements, respectively, for all the applied algorithms according to the three worry levels. The results indicate that although the dataset was balanced, the TFIDF feature extraction method did not provide sufficient information to the classifiers. The feature sets did not detect the worry classes due to embedded semantic features within this textual class label, which TFIDF could not capture.

The "no-worry" class was the highest-performing class label, while the "worry" class was the lowest-performing class label. However, for the "no-worry" and "high-worry" classes, other information was available. For example, in the "no-worry" class, some terms related to blame, news, and politics are present. As for the "high-worry" class, intensifiers were present. These results indicate the possible usefulness of the TFIDF feature set.

TABLE II. RESULTS OF THE PROPOSED MACHINE LEARNING (ML) MODELS

Algorithm	Accuracy (%)	Precision	Recall	F1-score
Multinomial NB	71	0.729	0.711	0.714
SVM	74	0.748	0.741	0.743
Logistic Regression	**75**	**0.751**	**0.747**	**0.748**
Random Forest	68	0.683	0.682	0.682

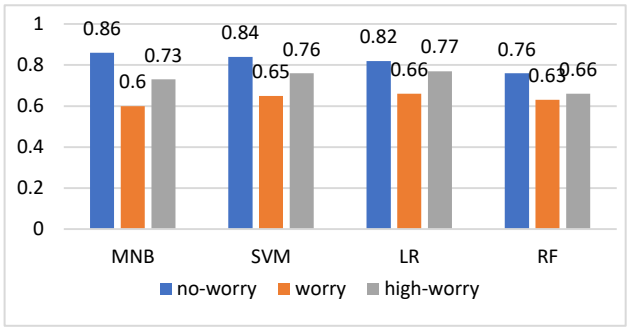


Fig. 5. Precision Results for the four ML Models per Worry Class.

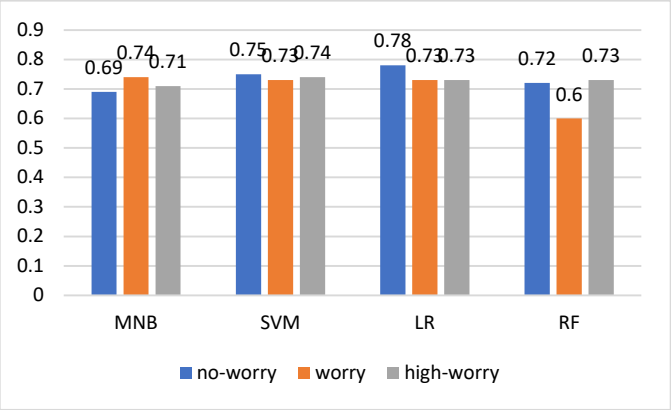


Fig. 6. Recall Results for the four ML Models per Worry Class.

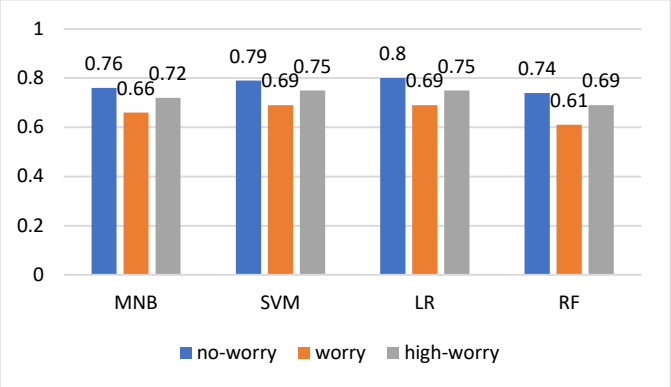


Fig. 7. F1-score Results for the Four ML Models per Worry Class.

In general, the current method, compared to Verma et al.'s study [14] is based on a new dataset. The new approach is also much more focused on the worry levels compared to 9 anxiety levels in [11], ranging from 1 to 9, where one was considered the lowest and nine the highest.

V. Conclusion

In this paper, we compiled a fine-grained benchmark dataset for the classification of worry levels concerning the COVID-19 pandemic. The dataset was collected from Twitter and was annotated using a majority vote among three experts. The WorryCov dataset was used to classify and predict the level of worry among Twitter users during the pandemic. Several experiments were conducted using the following ML algorithms: NB, LR, RF, and SVM. The optimal performance was achieved by LR, with an accuracy of 75%. It is recommended that the proposed approach be used for decision-making in healthcare entities to plan programs for the affected people. However, the current work has a few limitations. For example, the dataset is relatively small and was collected based on a short period from 2020–2021. Moreover, human behavior changes over time due to interaction with infected people, vaccination initiatives, and governments' health policies. Therefore, a new set of keywords that represent the new set of tweets might be needed to uncover the new trends in human worry levels. In the future, several deep learning models could be used to enhance the performance of the current approaches.

References

[1] C. Sohrabi et al., "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)," International Journal of Surgery, vol. 76. Elsevier Ltd, pp. 71–76, Apr. 01, 2020. doi: 10.1016/j.ijsu.2020.02.034.

[2] S. Li, Y. Wang, J. Xue, N. Zhao, and T. Zhu, "The impact of covid-19 epidemic declaration on psychological consequences: A study on active weibo users," Int J Environ Res Public Health, vol. 17, no. 6, Mar. 2020, doi: 10.3390/ijerph17062032.

[3] S. Zhang et al., "The COVID-19 Pandemic and Mental Health Concerns on Twitter in the United States," Health Data Science, vol. 2022, pp. 1–9, Feb. 2022, doi: 10.34133/2022/9758408.

[4] V. Sideropoulos, H. Kye, D. Dukes, A. C. Samson, O. Palikara, and J. van Herwegen, "Anxiety and Worries of Individuals with Down Syndrome During the COVID-19 Pandemic: A Comparative Study in the UK," J Autism Dev Disord, 2022, doi: 10.1007/s10803-022-05450-0.

[5] X. Ji, S. A. Chun, and J. Geller, "Monitoring public health concerns using twitter sentiment classifications," Proceedings - 2013 IEEE International Conference on Healthcare Informatics, ICHI 2013, no. September, pp. 335–344, 2013, doi: 10.1109/ICHI.2013.47.

[6] U. Qazi, M. Imran, and F. Ofli, "GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information," SIGSPATIAL Special, vol. 12, no. 1, pp. 6–15, Jun. 2020, doi: 10.1145/3404820.3404823.

[7] L. Sinnenberg, A. M. Buttenheim, K. Padrez, C. Mancheno, L. Ungar, and R. M. Merchant, "Twitter as a tool for health research: A systematic review," American Journal of Public Health, vol. 107, no. 1. American Public Health Association Inc., pp. e1–e8, Jan. 01, 2017. doi: 10.2105/AJPH.2016.303512.

[8] Q. Yang et al., "SenWave: Monitoring the Global Sentiments under the COVID-19 Pandemic," pp. 1–14, 2020.

[9] A. H. Alamoodi et al., "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review," Expert Systems with Applications, vol. 167. Elsevier Ltd, Apr. 01, 2021. doi: 10.1016/j.eswa.2020.114155.

[10] "Paul Ekman. 1999. Basic emotions. In".

[11] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets," IEEE Access, vol. 8, pp. 181074–181090, 2020, doi: 10.1109/ACCESS.2020.3027350.

[12] B. Kleinberg, I. van der Vegt, and M. Mozes, "Measuring Emotions in the COVID-19 Real World Worry Dataset," vol. 1, 2020.

[13] X. Li, M. Zhou, J. Wu, A. Yuan, F. Wu, and J. Li, "Analyzing COVID-19 on Online Social Media: Trends, Sentiments and Emotions," May 2020.

[14] R. Verma, C. von der Weth, J. Vachery, and M. Kankanhalli, "Identifying Worry in Twitter: Beyond Emotion Analysis," pp. 72–82, 2020, doi: 10.18653/v1/2020.nlpcss-1.9.

[15] R. A. Faisal, M. C. Jobe, O. Ahmed, and T. Sharker, "Replication analysis of the COVID-19 Worry Scale," Death Stud, vol. 46, no. 3, pp. 574–580, 2022.

[16] P. J. Schulz, E. M. Andersson, N. Bizzotto, and M. Norberg, "Using Ecological momentary assessment to study the development of COVid-19 worries in Sweden: Longitudinal study," J Med Internet Res, vol. 23, no. 11, p. e26743, 2021.

[17] B. Mandel, A. Culotta, J. Boulahanis, D. Stark, B. Lewis, and J. Rodrigue, "A demographic analysis of online sentiment during Hurricane Irene," Proceedings of the 2012 Workshop on Language in Social Media, no. Lsm, pp. 27–36, 2012.

[18] X. Ji, S. A. Chun, Z. Wei, and J. Geller, "Twitter sentiment classification for measuring public health concerns," Soc Netw Anal Min, vol. 5, no. 1, pp. 1–25, 2015, doi: 10.1007/s13278-015-0253-5.

[19] J. Pei, G. Yu, X. Tian, and M. R. Donnelley, "A new method for early detection of mass concern about public health issues," J Risk Res, vol. 20, no. 4, pp. 516–532, Apr. 2017, doi: 10.1080/13669877.2015.1100655.

[20] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hai, and Z. Shah, "Top concerns of tweeters during the COVID-19 pandemic: A surveillance study," J Med Internet Res, vol. 22, no. 4, Apr. 2020, doi: 10.2196/19016.

[21] Et al. Song M, Emilsson L, Bozorg SR, Nguyen LH, Joshi AD, Staller K, "Understanding Weekly COVID-19 Concerns through Dynamic Content-Specific LDA Topic Modeling," Lancet Gastroenterol Hepatol, vol. 5, no. 6, pp. 537–547, 2020, doi: 10.18653/v1/2020.nlpcss-1.21.Understanding.

[22] S. A. Chun, A. C. Y. Li, A. Toliyat, and J. Geller, "Tracking citizen's concerns during COVID-19 pandemic," in ACM International Conference Proceeding Series, Jun. 2020, pp. 322–323. doi: 10.1145/3396956.3397000.

[23] N. A. Hasanah, N. Suciati, and D. Purwitasari, "Identifying degree-of-concern on covid-19 topics with text classification of twitters," Register: Jurnal Ilmiah Teknologi Sistem Informasi, vol. 7, no. 1, pp. 50–62, 2021, doi: 10.26594/register.v7i1.2234.

[24] O. T. Aduragba, J. Yu, A. I. Cristea, and L. Shi, "Detecting Fine-Grained Emotions on Social Media during Major Disease Outbreaks: Health and Well-being before and during the COVID-19 Pandemic," AMIA Annu Symp Proc, vol. 2021, pp. 187–196, 2021.

[25] J. H. Shen and F. Rudzicz, "Detecting Anxiety through Reddit," Association for Computational Linguistics, 2017.

[26] D. Owen, J. C. Collados, and L. Espinosa-Anke, "Towards Preemptive Detection of Depression and Anxiety in Twitter," Nov. 2020.

[27] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets," NAACL HLT 2018 - International Workshop on Semantic Evaluation, SemEval 2018 - Proceedings of the 12th Workshop, pp. 1–17, 2018, doi: 10.18653/v1/s18-1001.

[28] J. Roesslein, "Tweepy: Twitter for Python!," URL: https://github.com/tweepy/tweepy, 2020.

[29] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," IEEE Trans Comput Soc Syst, vol. 8, no. 4, pp. 976–988, Aug. 2021, doi: 10.1109/TCSS.2021.3051189.

[30] Y. S. Malik et al., "Coronavirus Disease Pandemic (COVID-19): Challenges and a Global Perspective," Pathogens, vol. 9, no. 7, 2020, doi: 10.3390/pathogens9070519.

[31] M. Saif and S. Kiritchenko, "Understanding emotions: A dataset of tweets to study interactions between affect categories," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 2018, pp. 7–12.

[32] T. Sosea, C. Pham, A. Tekle, C. Caragea, and J. J. Li, "Emotion analysis and detection during COVID-19," ArXiv, 2021.

[33] S. M. Mohammad and F. Bravo-Marquez, "Emotion intensities in tweets," arXiv preprint arXiv:1708.03696, 2017.

[34] D. Ramyachitra and P. Manikandan, "Imbalanced dataset classification and solutions: a review," International Journal of Computing and Business Research (IJCBR), vol. 5, no. 4, pp. 1–29, 2014.

[35] S. Kusal, S. Patil, K. Kotecha, R. Aluvalu, and V. Varadarajan, "Ai based emotion detection for textual big data: Techniques and contribution," Big Data and Cognitive Computing, vol. 5, no. 3, 2021, doi: 10.3390/bdcc5030043.

[36] N. Babanejad, A. Agrawal, A. An, and M. Papagelis, "A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks," pp. 5799–5810, 2020, doi: 10.18653/v1/2020.acl-main.514.

[37] U. Naseem, I. Razzak, S. K. Khan, and M. Prasad, "A Comprehensive Survey on Word Representation Models: From Classical to State-of-the-Art Word Representation Language Models," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 20, no. 5, pp. 1–46, Jun. 2021, doi: 10.1145/3434237.

[38] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Information (Switzerland), vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[39] S. Adhikari et al., "Exploiting linguistic information from Nepali transcripts for early detection of Alzheimer's disease using natural language processing and machine learning techniques," International Journal of Human Computer Studies, vol. 160, Apr. 2022, doi: 10.1016/j.ijhcs.2021.102761.

[40] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," 2002. [Online]. Available: http://nltk.sf.net/.

[41] H. Du et al., "Twitter vs News: Concern analysis of the 2018 California wildfire event," Proceedings - International Computer Software and Applications Conference, vol. 2, pp. 207–212, 2019, doi: 10.1109/COMPSAC.2019.10208.

[42] Y. Didi, A. Walha, and A. Wali, "COVID-19 Tweets Classification Based on a Hybrid Word Embedding Method," Big Data and Cognitive Computing, vol. 6, no. 2, p. 58, 2022, doi: 10.3390/bdcc6020058.

[43] K. S. Kalaivani, S. Uma, and C. S. Kanimozhiselvi, "A Review on Feature Extraction Techniques for Sentiment Classification," Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020, no. Iccmc, pp. 679–683, 2020, doi: 10.1109/ICCMC48092.2020.ICCMC-000126.

[44] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," Procedia Comput Sci, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[45] O. Kramer, "Scikit-learn," in Machine learning for evolution strategies, Springer, 2016, pp. 45–53.

Local Pre-Conditioning and Quality Enhancement to Handle Different Data Complexities in Contactless Fingerprint Classification

Deepika K C[1]
Dept. of Electronics and Communication
Malnad College of Engineering
Hassan, Karnataka, India

G Shivakumar[2]
Dept. of Electronics and Instrumentation
Malnad College of Engineering
Hassan, Karnataka, India

***Abstract*—Biometric authentication systems have always been a fascinating approach to meet personalized security. Among the major existing solutions fingerprint-biometrics have gained widespread attention; yet, guaranteeing scalability and reliability over real-time demands remains a challenge. Despite innovations, the recent COVID-19 pandemic has capped the efficacy of the existing touch-based two-dimensional fingerprint detection models. Though, touchless fingerprint detection is considered as a viable alternative; yet, the real-time data complexities like non-linear textural patterns, dusts, non-uniform local conditions like illumination, contrast, orientation make it complex for realization. Moreover, the likelihood of ridge discontinuity and spatio-temporal texture damages can limit its efficacy. Considering these complexities, here, we focused on improving the input image intrinsic feature characteristics. More specifically, applied normalization, ridge orientation estimation, ridge frequency estimation, ridge masking and Gabor filtering over the input touchless fingerprint images. The proposed model mainly focusses on reducing FPR & EER by dividing the input image in to blocks and classify each input block as recoverable and nonrecoverable image block. Finally, an image with higher recoverable blocks with sufficiently large intrinsic features were considered for feature extraction and classification. The Proposed method outperforms when compared with the existing state of the art methods by achieving an accuracy of 94.72%, precision of 98.84%, recall of 97.716%, F-Measure 0.9827, specificity of 95.38% and a reduced EER of about 0.084.**

***Keywords*—Ridge orientation; Gabor filtering; region masking; ridge frequency; contactless fingerprint**

I. INTRODUCTION

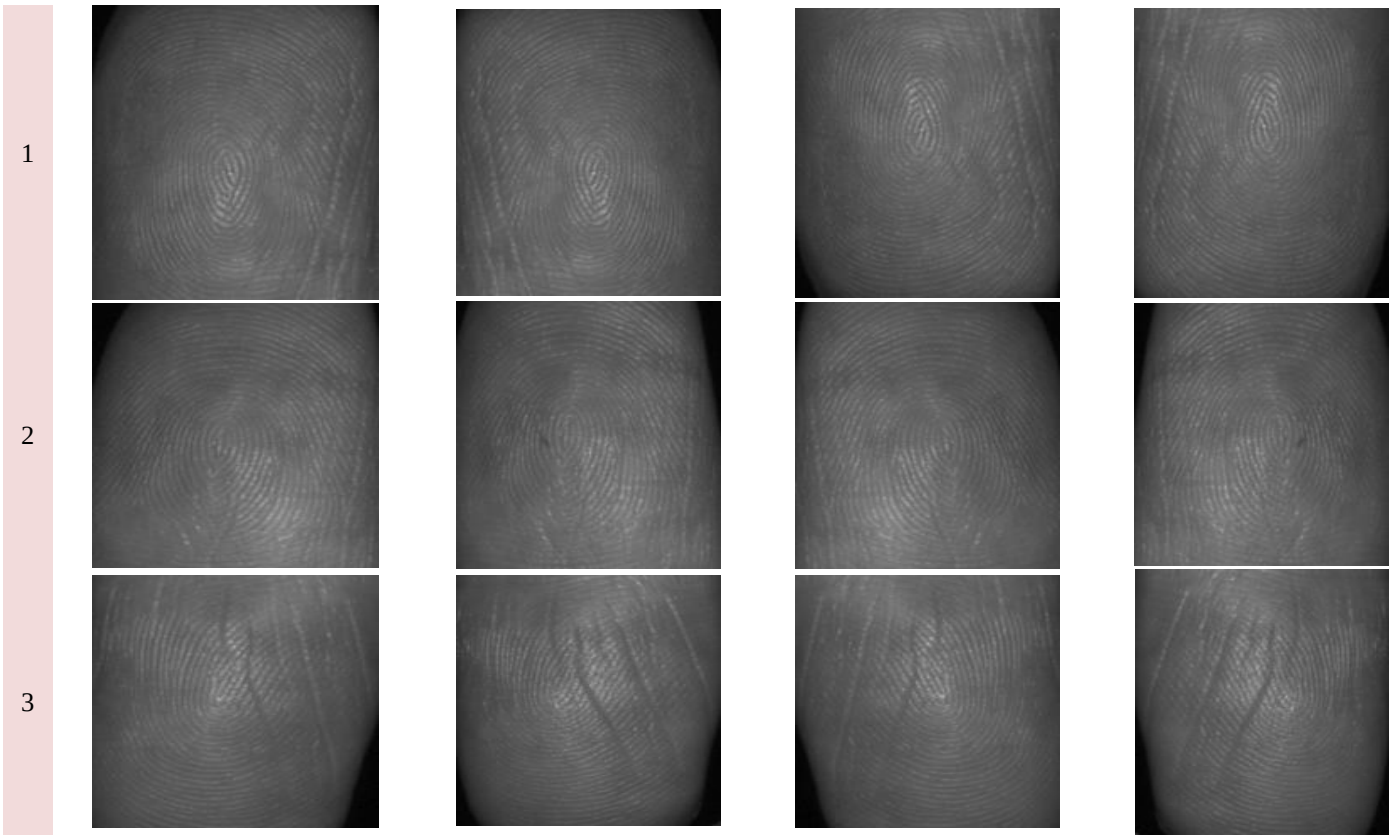
The last few decades have witnessed exponential rise in advanced technologies, including software computing, decentralized computing, smart intelligence, sensor and hardware systems. Despite significant innovation and technological horizon, personalized security or system security often remains a challenge under dynamic application environment [1]. Whether it is data, channel or infrastructure, guaranteeing security for these key systems has remained as an open challenge for academia-industries [2]. In the last few years, the rise in attack events too has increased significantly. The different attacks models have been developed on the basis of the exploiting user's or system access credentials like passwords, smart card attack loss, impersonation, Brute Force attacks etc. [1][2]. Most of these attacks have resulted huge data losses and breach, financial losses, system failure, and even the loss of life. Unlike cryptographic concepts, in the last few years biometric driven authentication systems have increased significantly [4][5] having superior potential with high scalability, interoperability and time-efficiency. Its efficacy can easily be visualized as Aadhar Card system by Unique Identification Authority of India (UIAI) [17]. Interestingly, more than a billion of population in India possesses a fingerprint driven Aadhar card for its verification. Though, Aadhar is a multi-modal system; however, evolved with fingerprint identification. In contemporary world whether it is corporate official attendance systems, entry or exit or even attendance systems in schools, fingerprint had remained a viable choice. In sync with such significances, a large number of efforts have been made by academia-industries; however, the recent pandemic of COVID-19 has limited the scope of the classical touch-based fingerprint authentication systems [6]. COVID-19 pandemic has almost limited the efficacy of the touch-based two-dimensional fingerprint driven modalities, as this pandemic was found exponentially spreading due to inter-personal infection through such frequently touching devices [14][15]. For instance, in certain offices, an executive could be seen trying his/her fingers many times to get system access. Fun, apart, but the severity of such frequent problems is high in real-world applications. The local conditions like sensor efficiency, optimality, sample distortion, scratches and humidity etc. often impact efficacy of the classical touch-based fingerprint techniques [3][7]. To alleviate such issues, improving feature modality in conjunction with contactless identification system seems to be the motivation for academia-industry for future efforts [8-12]. Noticeably, unlike touch-based two-dimensional feature learning environment, retrieving fingerprint feature under different orientation, lighting conditions is a complex problem. Moreover, suffer from the low accuracy and hence such system often undergoes false positive under varied local feature conditions and spatio-temporal complexities. Therefore, to cope up with touchless fingerprint identification system demands, researchers require improving local conditions, feature modalities as well as learning environment [13][16]. These key scopes are considered as the key driving forces behind this study.

In the last few years very few but significant efforts have been made towards touchless fingerprint detection methods. Jonietz et al. [3] recently tried to use depth camera and mobile

devices to perform touchless fingerprint detection. Despite RGB image and depth information combination, the key problem of non-uniform cues over feature space makes it limited towards realistic purposes. To alleviate such issues, Pang et al. [18] derived a three-dimensional feature model from input image that helped in improving ridge-valley information. To achieve it, authors at first employed least square method to fit a local paraboloid surface that helped estimating the local surface curvature and tensors curvature. Though, this approach helped in improving ridge-value orientation and depth information; however, at the cost of increased computation. Unlike previous works, Jonietz et al. [19] designed touchless finger detection model exploited aggregated channel features with RGB color space for finger segmentation that in conjunction with geometric shape helped estimating the fingertip for verification. Zaghetto et al. [20] too made effort to alleviate issues primarily caused due to orientational complexity and resulting spatio-temporal feature changes. To achieve it, authors applied Multiview scanner with multilayer neuro-computing. Despite their ability to address bad positioning problem, they could achieve the highest accuracy of 94%, which still needed to be improved. Though, Galbally et al. [8] made effort to improve accuracy by applying Laser sensing technique named 3D: FLARE. Yet, this approach was limited to yield a scalable solution for real-world purposes. Noticeably, these all approaches failed in providing a solution with scalability and efficacy towards real-world application. However, the depth assessment indicates that improving local input condition with superior feature segmentation and learning can yield superior performance.

Considering above stated key issues and allied scopes in this research, the emphasis was made on multi-dimensional optimization including pre-processing, feature extraction and eventual learning model. Being touchless approach, we considered normal three-dimensional RGB images as input, which is then processed for histogram equalization followed by contrast improvement and filtering. Recalling, non-linear ridge value patterns and local textural variations, we performed image normalization using Z-score method. Here, we performed block-wise normalization to improve contrast information. Subsequently, orientation image estimation was performed to improve local feature distribution. Moreover, it enabled frequency image estimation to make further spatio-temporal feature learning better. As post frequency image estimation, we performed ridge mask generation and Gabor filtering to ensure optimal local spatio-temporal textural feature (STTF) distribution for further minutiae detection. Unlike classical approaches, we performed three-dimensional minutiae projection and ridge mapping that improved overall feature space to achieve better spatio-temporal features for further learning. Finally, cropping the improved ridge mapping information, we performed deep-STTF feature extraction by applying Gray-level Co-occurrence Matrix (GLCM) followed by classification using random forest algorithm.

TABLE I. ILLUSTRATION OF 3D CONTACTLESS FINGERPRINT SAMPLES



II. METHODOLOGY

This section focuses on improving input data environment to ensure reliable fingerprint detection. In major touchless fingerprint detection models the viewing angle, image orientation, loss of ridges or damaged ridges and furrow structure, varying lighting or contrast etc. often impacts features, that eventually influences overall prediction accuracy. Considering this fact, in this paper, we focused on alleviating local data complexities. Moreover, we intend to guarantee intrinsic feature driven local conditioning so as to make optimal feature extraction without depending on the classical minutiae detection and segmentation. To achieve it, the proposed work encompasses data acquisition, Local Pre-conditioning and Image Quality Enhancement followed by feature extraction, classification and performance analysis.

A. Data Acquisition

In sync with the targeted contactless environment for fingerprint detection system, in this work we collected contactless three-dimensional sensor driven images to prepare datasets. The 3D touchless fingerprint datasets were collected in such a manner that it could enable effective learning under data heterogeneity and diversity to make it more efficient under realistic environment. Training over the large heterogeneous fingerprint patterns can make artificial intelligence driven models robust towards realistic purposes. Moreover, it can help achieving high reliability. We considered the 3D Fingerprint dataset comprising a large contactless fingerprint sample. Noticeably, for our case study we considered a total of 50 subjects and the samples collected were from the subjects aged in between 28 to 55 years. The subjects comprised a total of 40 man and 10 women that eventually contributed 160 and 40 fingerprint samples, correspondingly. The data considered had been collected under natural light conditions with standard illumination. Here, no specific light or illumination control measure was applied. To introduce diversity in reference to the viewing angle, illumination, contrast, orientation etc., subjects were instructed to stand in-front of the camera; and were instructed to move freely while keeping target fingers within camera vision range. Though, the similar dataset named 3D-FLRE-DB retrieves each fingerprint sample 15 times, where five different samples were obtained at a specific speed; we considered data retrieval at the random movement without any pre-calibrated speed definition. To introduce mode STTF feature heterogeneity the samples were not collected consecutively rather were captured at the different interval or gaps. To achieve it, once capturing one sample from a subject, the sample from another subject was captured, and this process was followed across the sample collection process over target subject volume. This approach was primarily done to introduce high spatial variability and textural heterogeneity to improve learning efficiency. A snippet of the data considered in this study is given in Table I.

1) Preliminary: Let, I be the input fingerprint image with $N \times N$ dimensional matrix, with $I(i,j)$ as the pixel intensity for the ith row and the jth column. In sync with touchless input, we hypothesize that the input images possess minimum resolution of 600 dots per inch, which is not difficult in contemporary high-definition camera. Thus, for the input images with aforesaid specification, the mean and the variance of the fingerprint image I in its gray-level form are derived as equation (1) and (2) respectively.

$$M(I) = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} I(i,j) \tag{1}$$

$$Var(I) = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \left(I(i,j) - M(I)\right)^2 \tag{2}$$

B. Local Pre-conditioning and Image Quality Enhancement

The overall proposed local pre-conditioning model encompasses the following key processes:

- Image Normalization.
- Local Orientation Estimation.
- Local Frequency Estimation.
- Ridge Masking.
- Gabor Filtering and Smoothening.

The proposed model at first performed normalization in such a manner that it retains a pre-defined mean and variance characterized.

2) Image normalization: Consider $I(i,j)$ be the gray-level value for the input touchless fingerprint image where (i,j) be the corresponding pixel values. Moreover, let M and Var be the measured mean and variance of the input image I. In this case, the normalized gray-level image for the input $I(i,j)$ can be obtained as $G(i,j)$, which is mathematically derived as per equation (3).

$$G(i,j) = \begin{cases} M_0 + \sqrt{\frac{Var_0(I(i,j)-M)^2}{Var}}, & if\ I(i,j) > M \\ M_0 - \sqrt{\frac{Var_0(I(i,j)-M)^2}{Var}}, & Otherwise \end{cases} \tag{3}$$

In (3), M_0 and Var_0 represents the expected mean and the variance values, correspondingly. In the proposed model, normalization is performed as a pixel-wise function and therefore it retained the native image clarity, especially ridge-and-furrow structure for further feature extraction and learning. Here, the key motive was to minimize the variations in the gray -level values in the direction of ridges and furrows so as to enable further processes more efficient without losing any intrinsic information.

3) Ridge orientation estimation: In reference to the touchless fingerprint, where the input image can have spatio-temporal differences caused because of varying light conditions, change in orientation, spatial and temporal feature non-linearity. To ensure optimal feature learning, we focused on improving ridge STTF. To achieve improved ridge information and allied intrinsic values, we performed ridge orientation estimation. In this work, we designed a least-mean square image orientation estimation concept for orientation image estimation. The proposed Ridge Orientation Image Estimation model is accomplished in multiple sequential steps. A snippet of the involved algorithm and allied implementation is given as follows:

Step-1: Split the input Gray-level image in $w \times w$ dimension. Here, we considered 64×64 dimension to split input image into multiple grids.

Step-2: Estimate the gradient information in x and y directions for each pixel element (i,j). Here, the gradient in x and y directions were, $\delta_x(i,j)$ and $\delta_y(i,j)$, respectively for the input pixel elements (i,j). In this work, to ensure low computational overheads over a large input image, we applied Sobel operator method to perform gradient estimation.

Step-3: Measure the local orientation values for each input block, especially centered at the pixel element (i,j) by applying following mathematical formula.

$$V_x(i,j) = \sum_{u=i-\frac{w}{2}}^{i+\frac{w}{2}} \sum_{v=j-\frac{w}{2}}^{j+\frac{w}{2}} 2\delta_x(u,v)\delta_y(u,v) \quad (4)$$

$$V_y(i,j) = \sum_{u=i-\frac{w}{2}}^{i+\frac{w}{2}} \sum_{v=j-\frac{w}{2}}^{j+\frac{w}{2}} \left(\delta_x^2(u,v) - \delta_y^2(u,v)\right) \quad (5)$$

$$\theta(i,j) = \frac{1}{2} tan^{-1}\left(\frac{V_y(i,j)}{V_x(i,j)}\right) \quad (6)$$

In (6), $\theta(i,j)$ represents the LMS value of the local ridge orientation for the block centred at the pixel location (i,j). In fact, ridge orientation signifies the direction which is orthogonal to the dominant direction of the Fourier spectrum over $w \times w$ window.

Step-4: This is the matter of fact that unlike touch-based fingerprint detection models, touchless image driven approaches might undergo more noise, reflections, dust related problems. In addition, touchless images can have the likelihood of the more damaged or corrupted ridge values or orientation, which can also be given rise due to the change in orientation or light intensity, contrast etc. Non-uniform skin surfaces too can show different spatio-temporal distribution for the ridge and furrow values in touchless fingerprint images. In sync with such complexities and allied challenges, the estimated values of the local ridge orientation can become inaccurate as well at certain time. In reference to these issues, we recall a hypothesis stating that as the local ridge orientation values vary gradually in local vicinity, especially in those neighboring localities where there is no singular point takes place or appear. In this reference, the use of a low-pass filter can be employed to manipulate the incorrect local ridge estimation (6). Now, to achieve it the orientation image is converted into a continuous vector field (CVF), which is mathematically derived as per (7) and (8). In above (7) and (8), the variables Φ_x and Φ_y represent the x and y components of the vector fields, correspondingly.

$$\Phi_x(x,y) = cos\big(2\theta(i,j)\big) \quad (7)$$

$$\Phi_y(x,y) = sin\big(2\theta(i,j)\big) \quad (8)$$

Then performed LPF filtering by applying following mathematical approaches (9-10).

$$\Phi'_x(i,j) = \sum_{u=-\frac{w_\Phi}{2}}^{\frac{w_\Phi}{2}} \sum_{v=-\frac{w_\Phi}{2}}^{\frac{w_\Phi}{2}} W(u,v)\Phi_x(i-uw, j-vw) \quad (9)$$

$$\Phi'_y(i,j) = \sum_{u=-\frac{w_\Phi}{2}}^{\frac{w_\Phi}{2}} \sum_{v=-\frac{w_\Phi}{2}}^{\frac{w_\Phi}{2}} W(u,v)\Phi_y(i-uw, j-vw) \quad (10)$$

In (9-10), the parameter W represents the two-dimensional LPF possessing single integral where the size of the filter is considered as $w_\Phi \times w_\Phi$. We performed smoothing at the block level where the filter size was fixed as 5×5.

Step-5: Update the local ridge orientation at the pixel position (i,j) by using (11).

$$O(i,j) = \frac{1}{2} tan\left(\frac{\Phi'_y(i,j)}{\Phi'_x(i,j)}\right) \quad (11)$$

Thus, applying above stated approach of smoothening and allied orientation image estimation we obtained a uniformly oriented field image, which is consequently processed for frequency estimation.

4) Ridge frequency estimation: As stated in the previous sections, in case of touchless fingerprint images, especially when there are no minutiae in certain neighborhood, the gray-level values along ridges can be reconstructed as a sinusoidal wave. Noticeably, these sinusoidal-shaped waves are modelled towards the direction orthogonal to the local ridge orientation. Because of this reason, another key intrinsic feature from the input fingerprint images can be obtained in the form of local ridge frequency estimation. In other words, similar to the ridge orientation, ridge frequency can be modelled as an intrinsic feature for the touchless fingerprint images. In the proposed model, to estimate the ridge frequency information in a neighborhood we employed the pre-estimated measures like normalized image and the ridge orientation images. Let, G and O be the normalized and the orientation images, respectively. Then, with these values, we estimated ridge frequency using following sequential implementation approach.

Step-1: Split the input Gray-level image in 64×64 dimension.

Step-2: Estimate the orientation window with fixed size $l \times w$ (128×64) over each block, centred at the pixel information (i,j).

Step-3: In reference to the Step-2, estimate the $x-$signature $(X[0], X[1], \dots, X[l-1])$ of the ridges within the window, conditioned at:

$$X[k] = \frac{1}{w}\sum_{d=0}^{w-1} G(u,v), k = 0,1,\dots,l-1 \quad (12)$$

$$u = i + \left(d - \frac{w}{2}\right) cosO(i,j) + \left(k - \frac{l}{2}\right) sinO(i,j) \quad (13)$$

$$u = j + \left(d - \frac{w}{2}\right) sinO(i,j) + \left(\frac{l}{2} - k\right) cosO(i,j) \quad (14)$$

In case there exists no minutiae in the oriented window, the x-signature constitutes a discrete sinusoidal-shape wave, possessing the similar frequency as that of the ridges in oriented window. This as a result, enables estimation of the ridge frequency from x-signature. Consider that $T(i,j)$ be the mean pixel counts in between the two subsequent peaks in the x-signature, then the ridge frequency $\Omega(i,j)$ is measured as per (15).

$$\Omega(i,j) = \frac{1}{T(i,j)} \quad (15)$$

In case there is no consecutive peaks available in x-signature, then the frequency is assigned a fixed value -1 that helps in differentiating it from the genuine frequency values.

Step-4: In case, the fingerprint images are taken over a predefined and definite resolution, then the value of frequency of the ridges within certain vicinity remains within a definite range. In case of 600 dots per inch resolution (DPI) this range remains within the level of $\left[\frac{1}{3}, \frac{1}{25}\right]$. In this manner, in case the measured value of the frequency becomes higher than the above stated range, the frequency is assigned a value -1, signifying that no genuine frequency could be estimated or observed.

Step-5: In touchless fingerprint images and corresponding blocks where the minutiae or ridges are corrupted due to any local or personal regions, it doesn't constitute any well-structured sinusoidal wave. In this case, it becomes inevitable to interpolate those frequency values of those specific blocks from the frequency of the adjacent blocks possessing well-structured frequency. Here, we applied the following measures to perform interpolation, over each block centered at the pixel location (i,j).

$$\Omega'(i,j) = \begin{cases} \Omega(i,j) & if\ \Omega(i,j) \neq -1 \\ \frac{\sum_{u=w_{\Omega/2}}^{w_{\Omega/2}} \sum_{l=w_{\Omega/2}}^{w_{\Omega/2}} W_g(u,v)\mu(\Omega(i-uw.j-uw))}{\sum_{u=w_{\Omega/2}}^{w_{\Omega/2}} \sum_{l=w_{\Omega/2}}^{w_{\Omega/2}} W_g(u,v)\delta(\Omega(i-uw.j-uw)+1)} & Otherwise \end{cases} \quad (16)$$

Where, $\mu(x) = \begin{cases} 0, & if\ x \leq 0 \\ x, & Otherwise \end{cases}$

$$\delta(x) = \begin{cases} 0, & if\ x \leq 0 \\ 1, & Otherwise \end{cases} \quad (17)$$

In (16), W_g refers the discrete Gaussian kernel with mean and variance are assigned as 0 and 9, correspondingly. Here, the other components w_Ω be the size of kernels which was fixed at 7. In case there exists minimum one block possessing the frequency value of -1, then the value of Ω is swapped to Ω′, and the above stated process is repeated (Step-5).

Step-6: Considering the gradual change in the inter-ridge distance variation, the proposed model applies LPF to eliminate the outliers.

$$F(i,j) = \sum_{u=w_{\Omega/2}}^{w_{1/2}} \sum_{u=w_{\Omega/2}}^{w_{1/2}} W_t(u,v)\, \Omega'(i-uw, j-uw) \quad (18)$$

In (18), W_t represents the two-dimensional LPF with single integral, while $W_t = 7$ be the filter's size.

5) Ridge masking: As stated above, in real-time touchless fingerprint image a block or allied pixel can be either in non-recoverable or recoverable region. And therefore, classification of the blocks or pixels in above stated categories can be done on the basis of the wave's shape analysis. In this work, we employed three distinct features including amplitude (α),frequency (β) and variance (γ). Consider that, $X[1], X[2], \ldots, X[l]$ be the x-signature of a specific block centered at the pixel position (i,j), then the aforesaid three different features pertaining to that block are obtained as per the following approach.

Step-1: Assign the value of α as the mean height of the peak and the mean depth of the valley.

Step-2: Define β as $\frac{1}{T(i,j)}$, where $T(i,j)$ refers the number of pixels in between the two consecutive peaks (average value).

Step-3: Estimate the value of variance γ, as per (19).

$$\gamma = \frac{1}{l}\sum_{i=1}^{l}\left(X[i] - \left(\frac{1}{l}\sum_{i=1}^{l} X[i]\right)\right)^2 \quad (19)$$

Thus, applying this method we estimated a large number of three-dimensional patterns for each input image. Moreover, k-NN classifier was applied to classify each block of $w \times w$ dimension that classifies each input block as recoverable or non-recoverable so as to help identifying the most suitable set of feature blocks for feature extraction. In case a block was found recoverable, the corresponding region was estimated. In case, the fraction of the recoverable region was lower in comparison to a predefined threshold $(T_{Threshold} = 40)$, we dropped that image for further feature extraction and learning. Finally, an image with higher recoverable image with sufficiently large intrinsic features were considered for further feature extraction and learning, so as to improve fingerprint detection and classification. Here, we label the recoverable region $R(i,j)$ as 1, while non-recoverable region is labelled as 0. Now, once identifying the optimal set of intrinsically enriched images, we performed filtering to improve spatio-temporal feature distribution. The details of the filtering method applied is given in the subsequent section.

6) Gabor filtering and smoothening: This is the matter of fact that the structure of the parallel ridges in fingerprint image, especially possessing well-structured orientation and frequency can provide sufficiently large intrinsic information to drop irrelevant and noisy components. On the other hand, the sinusoid waves pertaining to the ridges too change gradually in the local fixed orientation. Because of this reason, a bandpass filter can be designed in such a manner that it would eliminate all unexpected or undesired noise components, while retaining the true ridge information for further learning. In reference to this scope, Gabor filter can be a viable solution as it possesses both orientation-selective characteristics as well as frequency-selective characteristics in both frequency as well as spatial domains. Considering this fact, we applied Gabor filter as the bandpass filter to eliminate noise components while preserving genuine ridge structures in fingerprint images. The Gabor filter can typically be presented as (20).

$$h(x,y:\phi,f) = exp\left\{\frac{1}{2}\left[\frac{(xcos\phi)^2}{\delta_x^2} + \frac{(ysin\phi)^2}{\delta_y^2}\right]\right\} cos(2\pi f cos\phi) \quad (20)$$

where ϕ refers the Gabor filter's orientation, while f represents the frequency of a sinusoidal wave. The components

δ_x and δ_y be the space constants pertaining to the Gaussian envelope towards x and y, correspondingly. Here, the modulation transfer function of the considered filter is stated as per (21).

$$H(u,v:\phi,f) = 2\pi\delta_x\delta_y\, exp\left\{-\frac{1}{2}\left[\frac{[(u-2\pi/f)sin\phi]^2}{\delta_u^2}+\frac{(ucos\phi)^2}{\delta_v^2}\right]\right\} + 2\pi\delta_x\delta_y\, exp\left\{-\frac{1}{2}\left[\frac{[(u-2\pi/f)sin\phi]^2}{\delta_u^2}+\frac{(ucos\phi)^2}{\delta_v^2}\right]\right\} \quad (21)$$

In (21), $\delta_u = 1/2\pi\delta_x$ and $\delta_v = 1/2\pi\delta_y$.

To implement Gabor filters over each input touchless fingerprint image, three different parameters including the frequency of the sinusoidal wave u_0, filter orientation and standard deviation of the Gaussian envelope in the different directions δ_x and δ_y are considered. Here, the frequency characteristics of the filter f is estimated by employing local ridge frequency and the ridge orientation values. In the proposed model, the selection of trade-off between δ_x and δ_y is maintained in such a manner that higher the trade-off, more noise tolerant. However, it might cause spurious ridge information. On the contrary, smaller the values, the lower the Gaussian envelope, δ_x and δ_y. However, it might be less effective towards noise elimination. In this work, δ_x and δ_y values were assigned as 4.0, each. Now, consider that the input gray-level input fingerprint image be G, O be the ridge orientation image, while F be the ridge frequency image, and R be the recoverable mask. Then, the improved fingerprint image ε is obtained using the following equation.

$$\varepsilon(i,j) = \begin{cases} 255, & if\ R(i,j)=0 \\ \sum_{u=-w_g/2}^{w_g/2}\sum_{v=-w_g/2}^{w_g/2} h(u,v:O(i,j),F(i,j)G(i-u,j-v), & Otherwise \end{cases} \quad (22)$$

Thus, the final local pre-conditioned and improved fingerprint images are processed further for the feature extraction and identification.

C. GLCM Driven STTF Textural Features Extraction and Classification

In this research work, GLCM functions as a descriptive statistical feature distribution model assessing the probability of the pixel's gray scale values over an input fingerprint image. Functionally, it extracts high-dimensional statistical features. In this work, the varied STTF features are distributed uniformly throughout the pre-processed input image. In this work, over each input fingerprint image we extracted the different STTF features, which were later combined together to yield a composite feature vector for learning and classification. In this method, the retrieved spatio-temporal textural features were derived in the form of a matrix representing pixel intensities $I(x,y)$, centered on the pixels (x,y). In this manner, we extracted different spatio-temporal textural features for each input pre-processed images, with distinct probability matrix $P_{i,j}$. Here, the above stated probability matrix signifies the differences of the intensity between the pixel elements i and j that later helps in detecting motion patterns. In GLCM gray-scale refers the pair association along a direction, and therefore retrieving the gray-scale values can yield a matrix representing the association matrix among the pixels towards the target direction. We obtained symmetric matrix S by amalgamating the gray-scale information along with the allied transpose values. It enables estimation of the cumulative relationship among pixels in one direction. We normalized the symmetric association matrix S using (23) to obtain the probability matrix $P_{i,j}$.

$$P_{i,j} = \frac{S_{i,j}}{\sum_{i,j=0}^{N-1} S_{i,j}} \quad (23)$$

With the extracted values of $P_{i,j}$, the different STTF features including Contrast, Energy, Homogeneity, Correlation, Mean, Standard deviation, Variance, Kurtosis and Skewness are obtained. As stated, a total of nine STTF features were obtained for further feature learning. Here, our predominant goal was to retain maximum possible and significant features for learning and classification so as to achieve higher accuracy.

Once extracting above stated nine different GLCM features, we performed horizontal concatenation to estimate a composite feature vector for further learning. The composite GLCM feature obtained is given in equation (24).

$$GLCM_{Feat} = Conc\begin{pmatrix} CONT, ENE, HOM, CORR, \\ Mean, Var, STD, Kur, Skw \end{pmatrix} \quad (24)$$

Now, once estimating the composite feature vector ($GLCM_{Feat}$), we projected it for feature learning and classification. As stated, in this work we intended to exploit maximum possible feature instances to ensure optimal learning by Random Forest learning method and hence classification accuracy.

III. Results and Discussion

As stated above, in this section we mainly focus on assessing efficacy of the proposed contactless fingerprint detection and classification model, qualitatively as well as quantitatively. In other words, here we examine whether the use of local pre-conditioned image improvement yields superior performance. Before discussing the simulation results quantitatively, a snippet of pre-conditioned and enhanced results is given as follows.

Fig. 1(a) presents a random input 3D touchless fingerprint image. Here, it can easily be visualized that the illumination at the image center and bottom is relatively higher in comparison to the top corners. Moreover, the ridge structures in lower right bottom are unclear with high level of ambiguity. Furthermore, the straight division lines on the left side (bottom to top) can easily be visualized in this sample image, which can disrupt the ridge continuity to make further feature segmentation or allied feature learning. Noticeably, there are numerous local conditions such as low temperature, salty water contact by which the ridge values get changed temporarily. Though, with touch-based classical methods while pressing finger over the sensor, such local deformations get suppressed; however, in touchless fingerprint detection it can have decisive impact on feature learning and hence classification. To alleviate such issues, we performed local pre-conditioning to improve the ridge quality for further feature extraction. As repeatedly stated in the previous sections, we intended to guarantee ridge continuity over the different local conditions while ensuring that the ridges contain sufficient intrinsic features. To achieve

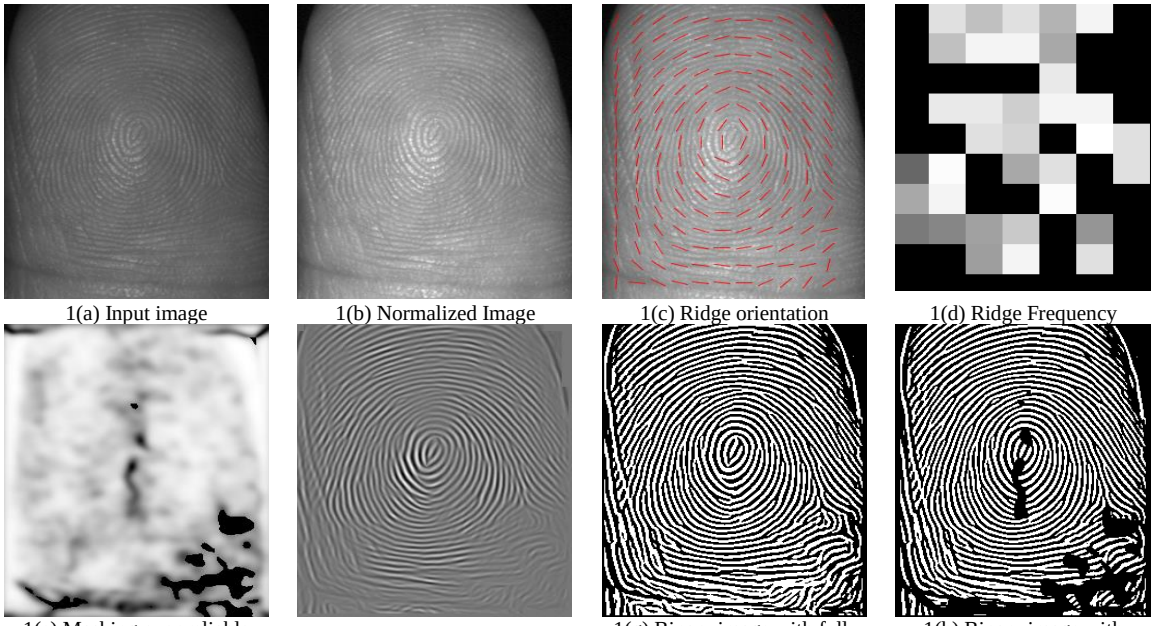
it, we applied the different pre-processing steps like image normalization, ridge orientation estimation, frequency estimation, ridge masking and filtering. Fig. 1(b) presents the normalized image output obtained from the original input image. Here, the impact of normalization can easily be visualized. Now, recalling the methodological intend where we intended to improve ridge structure continuity even over non-linear textural fingerprint surfaces, we performed ridge orientation estimation as shown in Fig. 1(c). The ridge frequency obtained over each grid is given in Fig. 1(d). In Fig. 1(e) presents the ridge masking results where the high frequent ridges are masked as 1, while the less frequent ridges are labelled as 0. Here, the key motive was to retain the ridge information carrying densely distributed features. The improved ridge structure is obtained by filtering (Fig. 1(f)). Here, observing the results it can easily be understood that the improved 3D touchless fingerprint image caries more uniform ridge's distribution with precisely perceptible structure, which can provide more efficient feature vectors for further learning and classification. The other images (Fig. 1(g) and Fig. 1(h)) represent the binary images, where 1(g) depicts the binarized image over the input (1(f)). Observing the bottom of the binarized image (Fig. 1(g)), it can be found that the bottom of the image carries ambiguities primarily because of ridge and furrow diversity, conjunction and non-linear bifurcation, random cuts etc. This as a result can impact STTF textural features and hence overall fingerprint detection accuracy. However, retaining a threshold driven approach can retain only feature intensive components to perform further feature and classification. Thus, observing the results in Fig. 1, it can be stated that the inclusion of the proposed model can yield superior feature vector for further learning and classification. Noticeably, in our proposed model to perform feature extraction we considered the improved ridge image (Fig. 1(f)) as input, which is hypothesized to yield superior performance.

The statistical performance outputs were measured by obtaining confusion matrix in terms of Accuracy, Precision, F-measure, Specificity, Recall and EER and are listed in Table II.

This is the matter of fact that a large number of studies have been done towards touch-based fingerprint detection systems; however, the efforts made towards touchless fingerprint detection are countable and very rare. Our depth literature assessment revealed that merely countable a dozen of efforts is made so far to introduce 3D touchless data for fingerprint detection. To assess relative performance, we have selected the recent methods like [8-12]. Ritesh and Ajay [9] developed a collaborative paradigm by exploiting ridge-valley minutiae information to perform contactless fingerprint detection. In their effort, authors mainly focused on improving minutiae under complex input data environment (like unclear ridge bifurcation, varied viewing angle and allied textural gradiance). Moreover, authors tried to suppress spurious minutiae information so as to improve accuracy and reliability.

TABLE II. PERFORMANCE ASSESSMENT

	Proposed method
Accuracy	94.72%
Precision	98.84%
Recall	97.71%
Specificity	95.38%
F-measure	0.9827
Equal Error Rate	0.084



1(a) Input image 1(b) Normalized Image 1(c) Ridge orientation 1(d) Ridge Frequency

1(e) Masking over reliable feature space 1(f) Improved Ridge structure 1(g) Binary image with full-scale input 1(h) Binary image with Threshold (Reliability >0.5)

Fig. 1. Pre-conditioned and Enhanced Results by the Proposed Method.

Authors applied the different datasets or fingerprint matchers like NBIS Biometric Image Software, MCC, and COTS (Commercial off-the-shelf), and therefore obtained the different performance over the different benchmark data. Interestingly, over the NBIS matcher they could achieve the EER of 13.33%. Noticeably, in comparison to their effort our proposed model exhibited EER of 0.084%. It shows superiority of our proposed model over the existing approach [9]. Recently, Galbally et al. [8] developed 3D-FLARE, a touchless fingerprint detection model; however. Despite the fact that their approach was quite complex in real-world realization, it exhibited EER of 10.03%. Though, to alleviate aforesaid data environment complexities, authors [8] made effort to segment yaw angle with fingerprint and fingertip separation etc., which was followed by hybrid feature extraction using local binary patterns (LBP) and Histograms of Oriented Gradient (HOG) features. Authors applied LBP+HOG features obtained from the segmented features to perform classification.

Authors could achieve the average EER of 10.03%, which is still higher than our proposed model. Kumar and Kwong [10] proposed a single camera driven touchless fingerprint detection model. In fact, it was a 3D minutia matching concept that made effort to recover extended 3D fingerprint features from the reconstructed 3D fingerprints. The EER performance by authors [10] was 1.02%, which is far more than our proposed model. An improved model by Lin and Kumar [11] applied deep learning driven multi-view touchless fingerprint detection model. This approach exploited multi-view deep representation to perform touchless fingerprint detection. Their proposed model [11] encompassed convolutional neural network where one fully convolutional network was applied to perform fingerprint segmentation, while three other layers were employed to learn 3D multi-view fingerprint feature representation. Undeniably, authors made effort to address at hand complexities with contactless fingerprint detection models that resulted into reduced EER value (0.64%). Zheng and Kumar [12] performed 3D fingerprint identification by exploiting recovered surface normal and albedo information. The key novelty of this approach was that it didn't require any surface reconstruction rather it employed different mathematical approaches to retrieve surface normal and albedo information, which was later used for learning and classification. The EER performed by this approach was 2.49%, which was higher than our proposed model. Thus, observing overall performance outcomes and allied inferences as shown in Fig. 2, it can be stated that the proposed touchless fingerprint detection model outperforms other state-of-the-art methods.

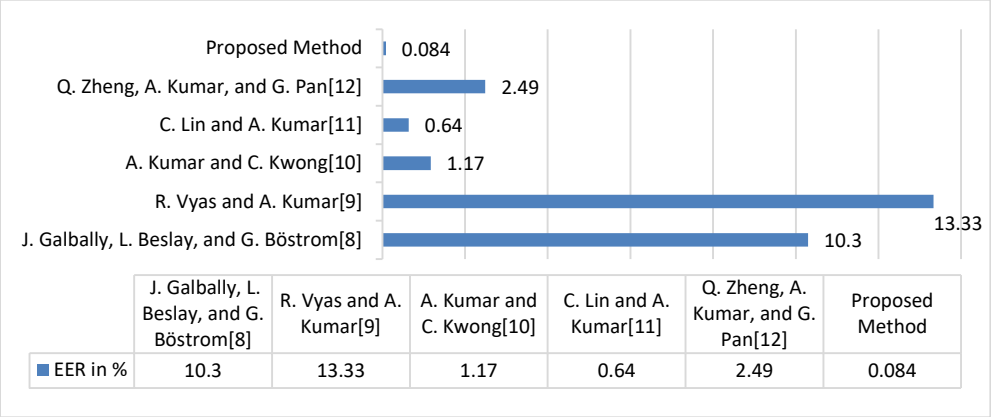


Fig. 2. Comparison of Equal Error rate of the Proposed Method with the Existing State-of-the-Art Methods.

IV. Conclusion

Since the inception, the fingerprint detection models have always been considered as a vital alternative of the classical cryptosystems. Undeniably, being fast in execution and diverse in spatio-temporal presentation, fingerprint-based systems turn out to be more efficient solution for personalized security and access control purposes. This efficacy makes fingerprint-based authentication system as one of the most used approaches for corporates, financial sectors, smart home and industrial monitoring and control. Despite robustness, being touch-based paradigm, its optimality has been challenges under different operating environment, especially in reference to the health and hygiene. During the recent pandemic of COVID-19, touch-based fingerprint models were found vulnerable due to touch-based infection probability. To alleviate such issues, contactless fingerprint detection method can be a viable solution; however, being touchless in nature such approaches might undergo different complexities like the impact of viewing angle, textural non-linearity, non-uniform illumination and contrast, ridge and furrow ambiguity, ridge discontinuity, etc. On the other hand, extracting structural features or other STTF features over aforesaid local adversaries can impact overall efficacy. In other words, training over a feature obtained from ambiguous or minimally distinct spatio-temporal feature space can give rise to the high false positive rate (FPR) and Equal Error Rate (EER). To alleviate such problems, it requires multiple optimization measures including local quality improvement or ridge improvement, and information-rich feature extraction. To achieve it, at first a local pre-conditioning concept was derived that mainly focused on improving ridge's orientation and spatial presentation so that the optimal features could be extracted. Recalling the fact that extracting features over the ambiguous ridges or furrows or even over detached ridges can lead false positive, the proposed pre-processing model helped in alleviating aforesaid complexities. This approach eventually retains only those feature-rich spatial components having clearly observable or distinctly distributed ridges for reliable feature extraction and

classification. As a future work we can experiment with the different feature extraction methods and learning algorithms to improve the accuracy of classification. Efforts can also be made in feature extraction stage like using Deep Neural Networks to reduce the Equal Error Rate and False Positive rate.

ACKNOWLEDGMENT

The Authors would like to thank the management, Principal and authorities of Malnad College of Engineering, Hassan for extending full support in carrying out this research work.

REFERENCES

[1] V. A. Thakor, M. A. Razzaque and M. R. A. Khandaker, "Lightweight Cryptography Algorithms for Resource-Constrained IoT Devices: A Review, Comparison and Research Opportunities," in IEEE Access, vol. 9, pp. 28177-28193, 2021, doi: 10.1109/ACCESS.2021.3052867.

[2] Krishna, P. G., & Muthuluru, N. Feistel network assisted dynamic keying based SPN lightweight encryption for IoT security. International Journal of Advanced Computer Science and Applications, 12(6). doi.org/10.14569/IJACSA.2021.0120642.

[3] C. Jonietz and I. Jivet, "Touchless Fingerprint Capturing from RGB-D Images in Mobile Devices," 2018 International Symposium on Electronics and Telecommunications (ISETC), 2018, pp. 1-4, doi: 10.1109/ISETC.2018.8583879.

[4] S. Ding, W. Bian, H. Liao, T. Sun and Y. Xue, "Combining Gabor filtering and classification dictionaries learning for fingerprint enhancement," in IET Biometrics, vol. 6, no. 6, pp. 438-447, 11 2017. https://doi.org/10.1049/iet-bmt.2016.0161.

[5] X. Si, J. Feng, J. Zhou and Y. Luo, "Detection and Rectification of Distorted Fingerprints," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pp. 555-568, 1 March 2015, doi: 10.1109/TPAMI.2014.2345403.

[6] N, Shanthi V, Aalelai Vendhan M. Rejuvenation of online research interactive fora during COVID-19. Indian Journal of Science and Technology .2020;13(47):4603605. https://doi.org/10.17485/IJST/v13i47.2230.

[7] Deepika, K.C., Shivakumar, G. A, "Robust Deep Features Enabled Touchless 3D-Fingerprint Classification System". SN Computer Science , volume 2, Article number: 263 (2021). https://doi.org/10.1007/s42979-021-00657-x.

[8] J. Galbally, L. Beslay and G. Böstrom, "3D-FLARE: A Touchless Full-3D Fingerprint Recognition System Based on Laser Sensing," in IEEE Access, vol. 8, pp. 145513-145534, 2020, doi: 10.1109/ACCESS.2020.3014796.

[9] R. Vyas and A. Kumar, "A Collaborative Approach using Ridge-Valley Minutiae for More Accurate Contactless Fingerprint Identification", Technical Report No.: COMP-K-25, 2018, pp. 1-15. https://doi.org/10.48550/arXiv.1909.06045.

[10] A. Kumar and C. Kwong, "Towards Contactless, Low-Cost and Accurate 3D Fingerprint Identification", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37, No. 3, March 2015, pp. 681-696. https://doi.org/10.1109/CVPR.2013.441.

[11] C. Lin and A. Kumar, "Contactless and Partial 3D Fingerprint Recognition using multi-view Deep Representation", Pattern Recognition (2018), DOI: 10.1016/j.patcog.2018.05.004.

[12] Q. Zheng, A. Kumar and G. Pan, "Contactless 3D fingerprint identification without 3D reconstruction," 2018 International Workshop on Biometrics and Forensics (IWBF), 2018, pp. 1-6, doi: 10.1109/IWBF.2018.8401566.

[13] K. C. Deepika and G. Shivakumar, "Hybrid CNN-Ensemble based Classifier for Touchless Fingerprint Classification," 2021 IEEE Mysore Sub Section International Conference (MysuruCon), 2021, pp. 482-486, https://doi:10.1109/MysuruCon52639.2021.9641718.

[14] Q. D. Vo and P. De, "A Survey of Fingerprint-Based Outdoor Localization," in IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 491-506, Firstquarter 2016, doi: 10.1109/COMST.2015.2448632.

[15] R. Kumar, P. Chandra and M. Hanmandlu, "Fingerprint Matching Using Rotational Invariant Image Based Descriptor and Machine Learning Techniques," 2013 6th International Conference on Emerging Trends in Engineering and Technology, 2013, pp. 13-18, doi: 10.1109/ICETET.2013.4.

[16] Deepika K.C., Shivakumar G. (2021) Towards More Accurate Touchless Fingerprint Classification Using Deep Learning and SVM. International Conference on Information Processing, Data Science and Computational Intelligence pp 248–257, 2021. https://doi.org/10.1007/978-3-030-91244-4_20.

[17] https://uidai.gov.in/.

[18] X. Pang, Z. Song and W. Xie, "Extracting Valley-Ridge Lines from Point-Cloud-Based 3D Fingerprint Models," in *IEEE Computer Graphics and Applications*, vol. 33, no. 4, pp. 73-81, July-Aug. 2013, doi: 10.1109/MCG.2012.128.

[19] C. Jonietz, E. Monari, H. Widak and C. Qu, "Towards mobile and touchless fingerprint verification," *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2015, pp. 1-6, doi: 10.1109/AVSS.2015.7301751.

[20] C. Zaghetto, B. Vidal and L. H. M. Aguiar, "Touchless multiview fingerprint quality assessment: rotational bad-positioning detection using Artificial Neural Networks," *2015 International Conference on Biometrics*, 2015, pp. 394-399, doi: 10.1109/ICB.2015.713910.

English and Arabic Chatbots: A Systematic Literature Review

Abeer S. Alsheddi
Computer Science Department
Imam Muhammad bin Saud Islamic University
King Saud University, Riyadh, Saudi Arabia

Lubna S. Alhenaki
Department of Computer Science, College of Computer and Information Sciences, Majmaah University
Al-Majmaah, 11952, Saudi Arabia

***Abstract*—In recent years, the availability of chatbot applications has increased substantially with the advancement of artificial intelligence techniques, and research efforts have been active in the English language, which presents state-of-the-art solutions. However, despite the popularity of the Arabic language, its research community is still in an immature stage. Therefore, the main objective of this systematic literature review is studying state-of-the-art research – for both the English and Arabic languages – to answer the proposed research questions regarding the development approaches, application domains, evaluation metrics, and development challenges of chatbot applications. The findings show that researchers have devoted more attention to the education domain using retrieval-based approaches while the generation-based approach has grown in popularity recently for providing new responses tasks. Whereas the hybrid approach for ranking multi-possible responses of combining both previous approaches shows a performance improvement. Besides, most metrics used to evaluate chatbot performance are human-based, followed by bilingual evaluation understudy and accuracy metrics. However, defining a common framework for evaluating chatbots remains a challenge. Finally, the open problems and future directions are highlighted to help in developing chatbots with minimal human interference to simulate natural conversations.**

***Keywords*—Chatbots; Arabic language; development approaches; domain applications; evaluation metrics**

I. Introduction

A chatbot is an example of a computer application based on artificial intelligence (AI) that aims to simulate human behavior by conducting a conversation with users using natural language data. The most well-known social applications, such as Telegram and Facebook Messenger, are supported by chatbots. Several organizational benefits of using chatbots include 24-hour availability, endless patience, increased sales, and reduced operational costs [1]. These benefits have led to an increasing demand for the development of chatbots using AI techniques. However, developing effective chatbots that can respond at the level of an actual human is challenging due to the requirement of understanding user inputs, generating appropriate responses, and perceiving the context of the conversation [2].

Over the past decade, a rapid development of chatbots based on English language has taken place in many application domains. In last two years, several *surveys* have been published about chatbots, mostly focused on the implementation approaches, such as those [3],[4],[5]. However , some of the previous research have been limited to specific domains, those researches reviewing the techniques, characteristics, and approaches used in the development used in the development of an intelligent tutoring chatbot applied to education [6], [7]. Moreover, the research in [8] examined previous articles which showed that the personalized learning framework of chatbots helped students improve in their studies. Although the surveys in [9],[10],[11],[12] were careful investigations, they have different aspects *than* this SLR. For example, the study in [11] used a different database and selection criteria. Also, the studies in [9] and [12] presented different research questions. In addition, the evaluation measures and challenges of implementation are not highlighted in [10].

Although developing a chatbot follows similar approaches regardless of which language is being used, for languages such as Arabic, chatbot implementation is challenged by the language's rich morphology, multiple dialects, and orthographic ambiguity [13][14]. However, according to findings obtained from this SLR, in the past three years, research about Arabic chatbots has substantially increased but still has insufficient resources such as available data sets, pretrained models, and tools [13]. Furthermore, to date, few surveys have been done about Arabic chatbots to identify the techniques, metrics, and data sets used. However, chatbots that existed till 2018 are used to process the data in the survey of [15] and did not address some of the same research questions investigated in this SLR [16]. In addition, the survey in [17] highlights one approach to develop a chatbot instead of covering all of the three approaches that will be covered in this SLR. Also, the study in [18] takes a different perspective on a number of applications involving the chatbot. Finally, the study in [19] investigates the characteristics of Arabic chatbots. Overall, the Arabic chatbot research community is still at an immature stage, so the English studies are included to present state-of-the-art solutions.

The objective of this SLR is to present a general overview of the English and Arabic developed chatbots by deeply investigating the current articles in this field. It addresses the domain applications, and approaches used to develop a chatbot and compare these addressed aspects to both languages, English and Arabic. The challenges and evaluation metrics also will be considered. Furthermore, after discussing the findings, the open research problems and future directions for research will be highlighted. The rest of this paper is organized as follows: Section II briefly overviews the chatbot technology. The methodology that follows in this SLR is provided in

Section III. The discussion of finding, open research problems and future research directions are presented in Section IV. Finally, Section V reports conclusion.

II. Chatbot Approaches and Techniques

A. Chatbot Approaches

The selected articles in this SLR generate their responses using different approaches. These approaches can be categorized based on the response generation into rule-based, corpus-based, and hybrid approaches [20]. Fig. 1 presents the approaches along with their common techniques in the selected articles.

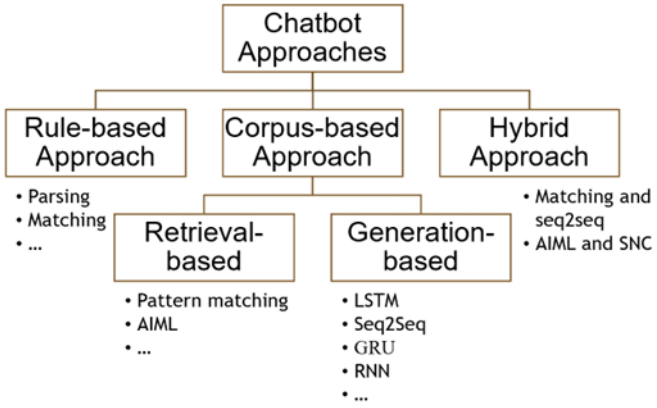


Fig. 1. The General Structure of Chatbot Approaches.

1) Rule-based approach is made up of a set of predefined human-made rules used in the hierarchy to convert user input into an output [20]. The rules break down the input into a sequence of tokens to find a pattern and generate a response rather than generating a new response. Although this approach can be considered easier in terms of implementing, it restricts responses to inputs *within* the predefined rules only and may provide inaccurate responses, leading to an unsatisfactory experience [1].

2) Corpus-based approach uses a knowledge base that contains a statistical language approach to select suitable responses instead of using predefined rules. Most chatbots in this approach produce their responses either two approaches:

- Retrieval-based approach uses information retrieval to get a candidate response from the corpus based on heuristics techniques rather than generating a new response [20]. This technique considers not just input but also the context by identifying keywords to offer the optimal response from a predefined responses (knowledge base) [21].
- Generation-based approach generates the responses based on the dialog context and does not require any predefined response. The chatbot attempts to generate a new response by considering the current and previous user interactions. It may require a large training set, which is potentially tricky to obtain. However, this approach has a high chance of response errors since generating in real-time.

3) Hybrid approach takes advantage of the combined strengths of the generation-based and retrieval-based approaches. Thus, the responses achieve more accurate results by capturing the informative pattern features. The researchers conducted *this* combination by ranking the response from generative and retrieval models or enhancing the informativeness and diversity retrieval responses by feeding them into a generation-based approach [1].

B. Chatbot Techniques

Several techniques are used in this SLR. The following subsections describe these techniques in detail:

1) Parsing. It converts text into meaningful representation string to determine the dependence relationship between its terms or its semantic structure. The parsing technique can be a lexical parsing that converts text into less complicated atomic terms to help extract information and simplify the manipulation. After applying lexical parsing, a syntactical parsing and a semantic parsing can be applied. These two parsing techniques determine a sentence's grammatical structure and extract a specific meaning by converting text to a machine-understandable representation of its meaning [21]. The parsing technique helps chatbots to understand text by identifying the main keywords in it [22]. For example, “set your eyes on my friend” and “could you see my friend” would both generate the same parsed “see my friend”. Moreover, this technique helps to identify the ambiguity in order to ask a user to rephrase his input [23]. For example, two possible ways to interpret the sentence “I saw my friend with my phone”: 1) did my phone help me see my friend; 2) did I see my friend holding my phone.

2) Pattern matching. Chatbots in this technique create a response with patterns where they are made manually, which is a non-trivial process [21]. Although it helps in response time reduction, the responses may be fully predictable and repeated, resulting in dull interactions that lack spontaneity and the human touch [24]. For example, the chatbot recognizes the input “where is the stickers?” and identifies keywords, here is “books”, where each keyword associates with an intent and response, here are “Office supplies” and “Different types of supplies in the office supplies aisle”, respectively.

3) AIML. Chatbots implement a syntax of the pattern matching technique through different technologies like the AIML to retrieve the most suitable response selection [25]. AIML is an open standard language derived from the Extensible Mark-up Language (XML). AIML comprises data objects consisting of two elements which are topics and categories. A topic is an optional top-level element with a set of categories related to that topic, while a category is a rule with a pattern and template matching for input and response, respectively. The objects are sorted in AIML files. Despite its readability, usefulness, and effective use of response time, it must provide a pattern for each conceivable response and update it on a regular basis, which cannot be done automatically [10].

4) RNNs. The RNNs allow the chatbot to handle sequential data and consider the current users' input. Due to the internal short memory, it memorizes the previous users' input. In other words, unlike a traditional Neural *Network*, RNN permits data to remain. The main idea of RNN is saving the output of a particular layer and feeding it as input to the next layer to predict the output [26]. Due to the vanishing or exploding gradient problem [27], the unmodified version of RNN is not appropriate for some applications. LSTM [28] and Gated Recurrent Units (GRU) [29],[30] are different solutions to this problem.

5) LSTM. The LSTM is a special kind of RNN [28]. LSTM is designed to handle long-term *dependencies* and solves the vanishing or exploding gradient problem in RNN. Thus, the gates are introduced in LSTM. Gates are the core component in LSTM, which decides which information that will be memorized. Additionally, the gates output the value between zero and one, where zero means do not memorize anything and one means let everything pass to the next state. Moreover, three kinds of gates are available in LSTM: input gates, forget gates, and output gates to control the flow of information. The input gate is responsible for the state update mechanism while the forget gates decide which information should be memorized. Additionally, the output gate determines the output from the hidden layer. The memory cell in LSTM comprises these three gates. Since the LSTM is created as the solution to short-term memory, it is capable of remembering aspects such as gender. Thus, depending on the previously remembered input, the chatbot can use "his/her". There exists a different architecture of LSTM, such as BiLSTM, which considers the input from the opposite direction as well [31]. Besides, the GRU is the main competitor of the LSTM and RNN [29],[30]. Due to its architecture, it is more popular and less complex than LSTM. The input and forget gate are combined to form a single "update gate".

6) Seq2Seq. Seq2Seq structure is the first architecture proposed to solve translation problems: their success bodes well for NLG. The seq2seq is trained end-to-end using different datasets and domains. Furthermore, due to its flexibility, simplicity, and generality, it is widely used to solve different NLP tasks, which makes seq2seq the industry-standard structure [32]. Technically, seq2seq is composed of two RNNs, namely, an Encoder and a Decoder. The Encoder processes the input of the user word-by-word while the Decoder generates the response word-by-word based on previous conversations. For building chatbots, rather than translating from one language to another, the problem was considered as translating the user input to the chatbot response. Additionally, the length of the input and response sequences can differ, which is one of the advantages of seq2seq structure over others.

III. Survey Methodology

The systematic method used in this SLR for reviewing chatbot articles are based on Kitchenham guideline [33]. This guideline consists of three stages which are planning, conducting, and reporting the review.

A. Planning

This subsection presents the research preparation of the SLR, search procedure and finally the inclusion and exclusion criteria.

1) Goals and research questions. The primary purpose of this SLR is focusing on analyzing the state-of-the-art English and Arabic chatbot articles, especially, with respect to their development approaches, application domains, evaluation metric and the main chatbot's development challenges. To achieve these objectives, four research questions are addressed. Table I presents the research questions and motivation.

TABLE I. List of the Research Questions

Number	***Research Question***	***Motivation***
RQ1	What are the main development approaches used for chatbot with regard to the user's generating appropriate response?	Identifying the state-of-the-art approaches and their techniques may be a significance for developers to provide more fit solutions by working and evolving the recent techniques trends
RQ2	What are the several domains used for building the chatbot?	Identifying the several application domains that most common used may help and enable researchers to address the current focus of domain of application as well as boost research in less contributed domains
RQ3	What are the commonly used metrics to evaluate the chatbots' performance?	Identifying the commonly used metric to evaluate the performance may help to improve and standardized the assess of chatbot, besides mitigate the difficult of comparing the performance of different chatbots
RQ4	What are the main challenges facing the implementation of Chatbot?	Open-research problem in development chatbot and the future directions are provided to continue developing the current issues.

2) Databases identification and search procedure. Six digital databases were *selected* which are: IEEE Xplore Digital Library, Springer, ScienceDirect, Google Scholar, Web of Science (ISI), and ACM. The search was done using 15 keywords as mentioned in [15], [9], [11] and presented in Fig. 2. The search procedure uses 15 keywords belonging to the computer science field in the predefined date ranges. There are 14 keywords for English and additional special keyword for Arabic search "Arabchat". For Arabic search, the same 14 keywords are used proceeding by adding "Arabic" term.

3) Inclusion and exclusion criteria. All criteria are applied manually after searching in the six databases. The inclusion criteria are based to the date and type of articles for both English and Arabic research. The selected articles of research work of the journals based on English chatbots started since 2018 till beginning 2022, period at which our actual SLR research study is done. However,due to the lack of Arabic research, the selected Arabic articles from journals and

conferences are starting from 2004 which is the publication year of one of the earliest Arabic articles till beginning 2022 [15]. Moreover, selected range of the earliest approach, rule-based approach, is extended because lately few recent research studies were utilizing this approach. Four inclusion criteria are: 1) Articles published in both English and Arabic Languages; 2) Articles published in specific ranges as mentioned previously; 3) Full text articles; 4) Articles that addressed the proposed research questions.

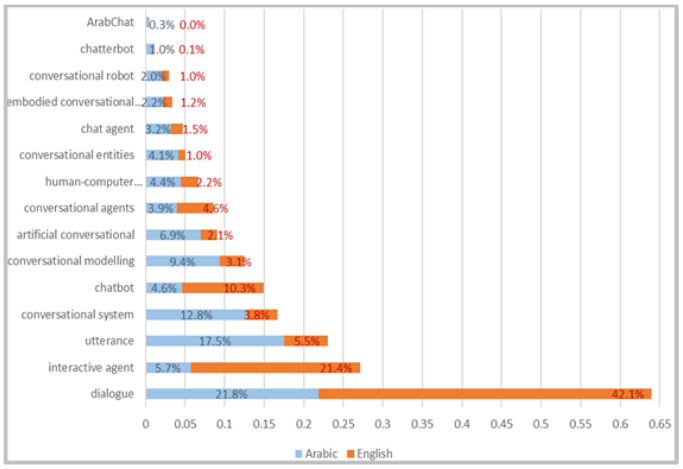


Fig. 2. Total Retrieved Articles for 15 Keywords.

After applying the inclusion criteria, four exclusion criteria are used. The first criterion discards any article that is not related to the chatbot depending on their titles and abstracts. Conspicuously the search in databases sometime returns a huge number of articles that cannot be processed manually. Thus, one ascension in this step is considered which is the returned articles in a database search engine are ordered relevantly to the keywords. Therefore, irrelevant investigated articles are followed by irrelevant articles as well. The second criterion is to remove the duplicate articles that appear in more than one database out of the searched six databases. The third criterion relates to the research questions and involves assessing the candidate articles under the quality assessment as will be discussed in the following subsection. The last criterion keeps journals for both languages and accepts Arabic articles from conferences, while books and theses are filters.

B. Conducting the Review

1) Article selection. The result of 15 keywords searching in all databases returns 59,169 articles. Fig. 2 presents the total number of returned articles for each keyword from all databases in ascending order. The most common keywords is “dialogue” for both English and Arabic chatbot research. Mostly, the greater number of words in each keyword, the smaller number of returned relevant articles, such as the two keywords "dialogue" and "human-computer conversational systems". Moreover, different searching strategies are used in digital databases for retrieving relevant articles. Some of them including in ACM and Google Scholar mostly return a large number of articles exceeding 250K which is caused to restrict the search in them to occur keywords only in the title of articles. Although of that, Google Scholar database still returns the largest number of relevant articles.

2) Data extraction. The result of applying manually exclusion criteria presents in Fig. 3. Firstly, removing irrelevant articles by analyzing their titles and abstracts remained 321 articles that were downloaded. Then 137 duplicated articles were removed. Next, the full texts of the remaining articles were investigated deeply, resulting therefore at filtering 56 articles not addressing the research questions or not passing the quality assessment. The 40 of articles are filtered involving books and theses. As a result, 50 and 38 English and Arabic articles, respectively, are relevant and investigated for the SLR.

3) Quality assessment. The assessment process was conducted in simultaneously with articles extraction. The process was performed for each candidate article individually where the various assessments is discussed until a consensus was achieved. The checklist of ten assessment questions are provided, where a candidate article was selected when at least it gets seven yes answered to the ten questions [34]: 1) Does the article present the objectives of the research clearly? 2) Does the article well-describe the proposed approaches and techniques? 3) Does the article attempt to address an existing issue with chatbot applications? 4) Does the design adhere to well-defined design concepts or principles? 5) Does the article describe the used dataset? 6) Does the article state the results? 7) Does the article state the process of performance evaluation? 8) Dose the article discusses one of domains in the chatbots application? 9) Does the article have a coherent reporting and understandable? 10) Does the article have an appropriate research method appropriate to address the aims of the research?

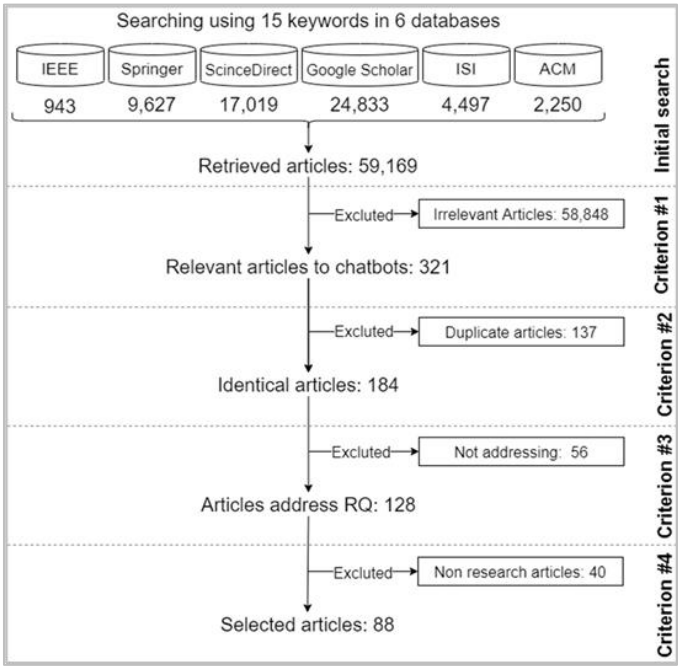


Fig. 3. Data Extraction Step.

C. Reporting the Review

1) Approaches. A detailed comparison of these works is offered in Table II. It is worth noting that no technique is the best selection for all problems. Hence, experimentations in the selected articles are used to compare these techniques. In addition, many chatbots have been developed using state-of-the-art platforms or programming languages for various purposes [5]. Table III presents publicly existing platforms in the selected articles to implement chatbots. The other research works done in [35],[36],[37],[38] develop chatbots using programming languages based on C# in Verbot 5.0, on Snatchbot, Microsoft Visual Studio linked to Google Translate API and the Twilio platform, respectively, as well as some libraries in python are used, such as ChatterBot and chatterbot_corpus.

2) Domains. The functionality of chatbots can be divided into two categories: task-oriented chatbots that interact to fulfill tasks, usually for a specific domain, and non-task-oriented chatbots that engage in open-domain interactions, usually for the sake of amusement. Depending on the selected articles, different domains are divided into two main categories: open domains and closed domains. Table IV analyzes English chatbots followed by Arabic chatbots ordered by their publication years from domain perspective.

3) Evaluation metrics. Human-based and automatic-based evaluation metrics are two main metrics to evaluate chatbots. Human evaluation involves having a group of people communicate with the chatbot and evaluating various aspects using evaluation frameworks or questionnaires. The other categories include different proposed evaluation frameworks, such as sensibleness and specificity average (SSA) and quasi-Turing test method [39][40]. Table V presents a summary and comparison of more than 20 types of evaluation metrics used in chatbot development over the selected articles. While more than 70 articles have detailed the evaluation of their works, about 15 different articles seem not to evaluate their chatbot at all, which were classified under the not available (NA) category. From the table, it can be observed the most common measure used for both languages is human evaluation, then the BLUE and Accuracy for English and Arabic, respectively.

TABLE II. APPROACHES AND TECHNIQUES USED IN IMPLEMENTING SELECTED ARTICLES

Approach	*Techniques/Structures*		*Description*	*Advantages*	*Disadvantages*	*Articles*
Rule-based	Rules and patterns matching		Set of predefined human-made rules	• Simple and lower cost implementation • Quick deploy • No overtime for learning user intent	• Cannot learn on their own. • Fail to response outside its preset understanding	[41],[42],[43]
	Parsing		Converting a text to be less complicated terms	• Providing dependency relationship between words or semantic structure of the text	• Same the rules and patterns matching disadvantages	[44],[45],[46]
Corpus-based: 1.Retrieval-based	Pattern matching		Matching inputs with predefined structures of responses	• Sufficient on simple tasks • Select informative responses from candidate responses • More flexible than the rule-based.	• Providing chatbots without reasoning and creation. • Limited capabilities and repeated responses.	[47],[48],[49],[50],[51], [52],[53],[54],[55],[56],[57], [58],[59],[60],[61],[62],[63], [64],[65],[66],[67]
	AIML		Represents the knowledge as objects which derived from XML	• Advantages of pattern matching • Powerful in designing conversational flow	• Building all the possible patterns manually • Difficult to scaling	[68],[69],[70],[71],[72],[73], [74],[75],[76],[77],[78]
2.Generation-based	LSTM	CNN	CNN usually used for learning features utomatically by utilizing convolution and pooling processes	• Accurate on a larger dataset • Suitable to remember longer sequences	• Uses large number of parameters • Required more memory size • Long execution time • More complex structure	[79],[80]
		CNN and GRU	GRU is a type of RNN technique related with LSTM.			[81]
		RGDDA based on gradient reinforcement learning	Generating responses by utilizing user-specific information			[40]
		Stacked LSTM	Stacked LSTM is comprised of multiple LSTM layers			[82]
		Stacked LSTM and BiLSTM	BiLSTM: Both directions is followed by the input.			[83]
		BiLSTM, HRED	HRED generates context and response			[84]

Approach	*Techniques/Structures*		*Description*	*Advantages*	*Disadvantages*	*Articles*
			embeddingusing three stacked RNNs			
		Attention	Attention is based on RNNs cell and improved technique of seq2seq learning			[85]
	seq2seq	Seq2seq learning based on encoder-decoder architecture	Mapping a sequence of input words to another representation of response sequence.	• Same advantages of tachniques that seq2seq depending on it • Support variable-length size of input and response	• Same disadvantages of tachniques that seq2seq depending on it	[39],[86],[87],[88],[89],[90], [91]
	GRU		In compression to LSTM, GRU required fewer parameters training and it not being required for an additional cell state	• Uses less training parameter • Uses less memory • Take less time in execution • Less complex structure	• Not suitable for large dataset • •Not suitable for long-distance relations	[92],[93]
	RNN	Attention mechanisms besides BiGRU	Attention mechanisms Improved technique of seq2seq learning based on RNNs cell. Resolve the problem of systems' incapability to remember a longer sequence	• Uses less training parameter • Uses less memory • Take less time in execution • Less complex structure	• Suffer from gradient exploding and vanishing problems. • Difficult to process very longer sequences • Not suitable for parallelizing or stacking up	[94]
		Enhancement of RNN-GRU	This technique is improved by adding three additional cells which are Refinement, Adjustment, and output cells			[95]
	Pre-trained model	GPT-2, DIALOGPT, BoB, aubmindlab, CakeChat, asafaya,	To map words to actual number vectors, a language modeling and feature extraction technique was used.			[96],[97],[98],[99], [100]
Hybrid	Combination of generation and retrieval-based approaches		XiaoIce: more popular example from Microsoft	• Easy to select the attributes (relevance) from ranked features list.	• If the hybridization technique is not complementary to each other, the performance quality may decrease.	[101]
			Proposed PS, GP, and PRF			[102]
			Develop matching method based on the seq2seq			[103]
			Develop a model using the Twitter LDA model and attention mechanism			[104]
			Integrate AIML technique with a SNC model			[105]
			Multi-strategy process including LSTM with an attention mechanism beside rule-based technique			[106]

TABLE III. PLATFORMS USED IN IMPLEMENTING SELECTED ARTICLES

Framework	*NLP features*	*API*	*Control conversation*	*Languages*	*Limitation*	*Articles*
Google's DialogFlow	Yes	Yes	Yes	More than 45 languages from Bengali to Vietnamese except Arabic	Limitation of understanding the synonyms and hyponyms besides the documentation isn't very good	[107],[108],[109],[110]
Pandorabots	Yes	Yes	No	Support many languages including Arabic	Limitations to dealing with Arabic spelling mistakes	[76],[77]
IBM Watson Conversation	Yes	Yes	NO	13 languages from Arabic to Spanish	Not support Enhanced intent detection and autocorrection fixes misspellings	[111],[112]
Microsoft Azure	Yes	Yes	Yes	Translating 100 languages from Afrikaans to Yucatec Maya	Replying using translation system which increase the error rate	[60]
Rasa	Yes	Yes	Yes	Can be trained on any languages	Not support Arabic predefined trained entities	[113],[114]
Facebook Bot Engine (Wit.ai)	Yes	Yes	Yes	More than 100 languages from Afrikaans to Zulu	Time consuming when training the chatbot to understand all the different forms of Arabic text	[115],[116],[117]
Chatfuel	NO	Yes	NO	Support many languages including Arabic	Inflexible in terms of conversation flows and multi-languages	[118]
OSCOVA	Yes	Yes	Yes	NA	Not appropriate for complex conversation	[119]
Recast.AI	Yes	Yes	Yes	More than 15 languages from Arabic to Swedish	Poor documentation	[120],[121]

TABLE IV. DOMAINS USED IN SELECTED ARTICLES

Domain	*Languages*	*Articles*
Religion	Classical Arabic	[73],[60],[72]
Education	English	[117],[86],[90],[108],[118],[107],[45],[112],[109],[122]
	Classical and MSA Arabic	[50],[51]
	MSA Arabic	[54],[58],[57],[55],[52],[53],[67],[66],[110],[63],[65]
	Arabic dialects: Saudi Arabic dialect and Jordanian	[77],[78]
Healthcare	English	[41],[44],[47],[70],[49],[71],[119]
	MSA Arabic	[74],[111]
	Arabic Dialects: Egyptian	[100]
Tourism and airline	English	[35],[43]
	MSA Arabic	[59],[75],[46],[113],[115]
Business and customer service	English	[87],[69],[121],[81],[84],[95],[83],[106],[120]
Empathy and personalization	English	[123],[101],[92],[124]
	MSA Arabic	[85],[125]
Open	English	[68],[89],[79],[40],[104],[99],[36],[102],[94],[39],[82],[48],[105],[88],[114],[98],[116]
	MSA Arabic	[64],[42],[61],[56],[80],[93],[38],[62]
	Arabic Dialects: Gulf Arabic and Egyptian	[91],[76]

TABLE V. METRICS USED IN SELECTED ARTICLES

Categorization	*Metrics*	*Articles*
Automatic based Metric	F1-Score	[89],[114],[106],[80],[62],[100]
	Precision	[105],[106],[80],[62]
	Recall	[105],[106],[80],[62]
	Accuracy	[92],[114],[126],[105],[124],[80],[93],[62],[100],[76],[56],[74]
	PPL	[89],[102],[94],[92],[98],[124],[126],[39],[101],[85]
	BLEU	[79],[83],[94],[81],[92],[95],[84],[126],[127],[86],[101],[49],[40],[40],[99],[85],[91]
	ROUGE	[79],[83],[84],[49]
	MAP, P@1 and MRR	[103],[104]
	SkipThoughts cosine similarity, embedding average cosine similarity, vector extrema cosine similarity, BOW and greedy matching scores	[79],[102],[94],[81],[84],[127],[88]
	Other	[87],[95],[84],[39],[116],[40],[99],[109],[43], [48], [79], [102], [127]
NA		[70],[41],[44], [107],[47], [68], [69], [108], [121],[38], [61], [67], [67], [73], [108], [113]
Human based Metric	H: User Satisfaction	[35], [45], [82], [84], [89], [118], [128],[36], [49], [98], [101], [117],[112], [119], [120], [122],[60], [77], [85], [90],[63], [78], [91], [110], [125],[46], [57], [59], [111], [115],[54], [55], [65], [66], [75],[42], [50]–[53], [58], [64], [72]

IV. Discussion

This SLR reviews 88 articles to address the four research questions. This section discusses findings, highlights challenges and open problems with providing future research directions that we expect would help in developing chatbots.

A. Finding

1) Approaches. Initially, no approach or technique is the best for all domain applications. Selecting one approach or another deeply depends on several considerations. With recent improvement in computational recourse, most articles use generation-based approach, which is increasingly important in the development of chatbots. The encoder-decoder architecture is used as the main learning method in chatbots. However, a large number of datasets and high amount of computation time is required. When it comes to selecting the appropriate response from the structured data to respond to the user's input, the retrieval-based approach is the best. There are different improvements for retrieved information in this approach as shown in [47],[65],[66],[69]. However, purely retrieval-based approaches do not perform reasoning, and therefore, they are only suitable for mirroring current knowledge. More recently, significant findings based on performance has been offered by the emergence of the various hybrid approaches. This improvement may be due to the advantages of combining previously mentioned approaches; retrieval-based and generation-based. Moreover, the rule-based approach is straightforward, easy to implement, understand and fast but is too fixed to predefined rules in the database. Thus, extraneous inputs cannot be answered. Thus, it is limited used in developing chatbots in comparison to other approaches.

Furthermore, the generation-based approach in English-language chatbots has more attention. However, overall, there is currently limited research on Arabic-language chatbots. Retrieval-based approach represents the vast majority of the selected approaches in Arabic chatbot research articles, whereas generation-based approach has started to attract attention in 2018 [80]. The most commonly used technique for Arabic-language chatbots is the pattern matching followed by AIML in retrieval-based approach. The LSTM technique is the most employed in developing Arabic chatbots followed by seq2seq and GRU in the generation-based approach.

Besides the three main approaches, some selected articles use publicly available platforms to create and launch their chatbots as presented in Table III. These platforms eliminate the required experience in coding to develop the chatbots from scratch. The platforms simplify development and standardize some implementation processes. Although some of these platforms have a well-described approach, such as Pandora's [129] that uses the retrieval-based approach, most hide the details of their systems. In addition, A closer inspection of the table shows that Google's DialogFlow, Facebook's Wit.ai, IBM Watson Conversation, and Microsoft Azure are cloud-based platforms that support different programming languages besides the natural languages. However, they differ significantly in other aspects [130].

2) Domains. The selected articles show most efforts was devoted to the education domain. A possible explanation might be because of several possible potential areas of education where chatbots can be utilized. According to Fig. 4, there is a significant difference in the published research between Arabic and English chatbots, especially in the business domain. Simultaneously, some of the Arabic chatbots are developed as religious chatbots, whereas English chatbots focus on other domains. Overall, the remaining domains are addressed almost equally in both languages. Furthermore, the research to date has tended to focus on MSA rather than Arabic dialects. This finding may be explained by the fact that MSA has a formal and clear format in written and expression, which helps with analyzing. However, due to the rise of social media platforms, few recently published articles focused on Arabic dialect chatbots, such as those in the Gulf Arabic dialect, Saudi dialect, Egyptian dialect and Jordanian dialect.

In addition, a relationship is observed between the domains and the approaches. Fig. 5 displays the distribution of the investigated approaches in 88 articles across seven domains. Obviously, for the education and health domains, the majority of chatbots seem to be developing using the retrieval-based approach. The possible explanation of this bias is that the chatbots in these domains are always prepared to fit a specific knowledge base. In contrast, most of the chatbots in business, emotions, and open domains are built using the generation-based approach. Thus, chatbots should be able to produce a new natural conversation with more appropriate responses.

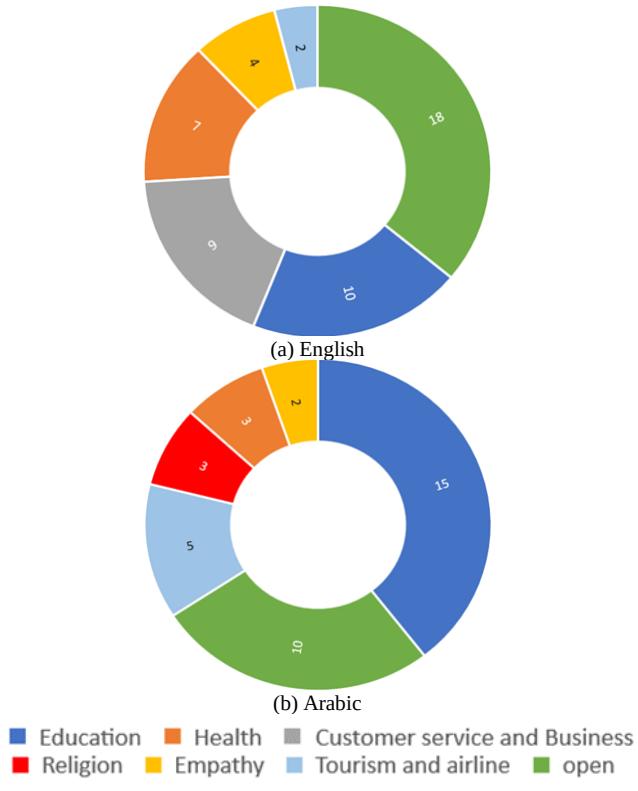


Fig. 4. Finding in Domains with Languages Perspective.

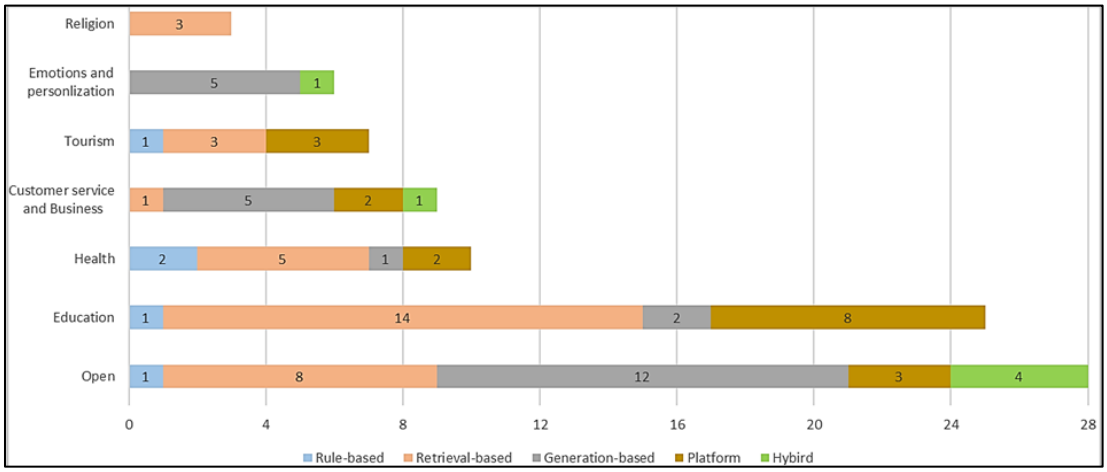


Fig. 5. Finding in Domains with Approaches Perspective.

3) Evaluation metrics. The vast majority of chatbot research uses the human-based metric, followed by BLUE and then accuracy, as presented in Fig. 6, where the most used types of the ROUGE metric are ROUGE-1 (unigram) and ROUGE-2 (bi-gram). Overall, no specific evaluation metric is more represented in articles due to the lack of gold standards of the evaluation. The evaluation of empathy, user satisfaction, and fluency are examples of the needed intervention of human evaluation. In terms of the time and resources, automated evaluation measures are more efficient than human ones. However, they appear to be incapable of accurately assessing the quality and efficacy of the entire conversation even they are easier to use and do not require manual work by human judges.

Moreover, several important differences are between the Arabic and English metrics. From Fig. 6, most of the Arabic articles focused on human evaluation whereas human evaluation and the BLEU metric are more popular in English articles. This may be due to several challenges of the Arabic language such as the lack of available data resources.

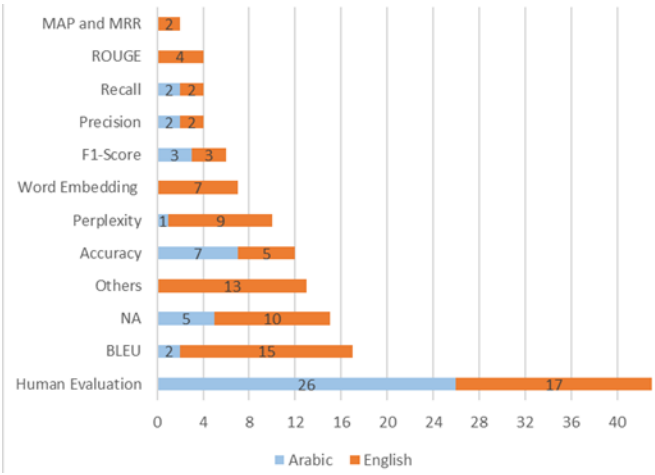


Fig. 6. Finding in Metrics Perspective.

B. Open Research Problems

Several challenges are in chatbots development, but for a comprehensive conversation experience, the development requires a balance between different directions.

1) Ambiguity in conversation. The simple meaning in natural language can be interpreted in more than one way. Therefore, simulating human conversation with a chatbot is quite challenging. Although deep learning has recently taken a lot of researcher attention, there is still a long way to go before chatbots can produce human-like conversation. Moreover, ambiguous word senses are a problem in the building of a semantic conversational AI.

2) Consistency in natural language interpretation. Inaccurate and unsuitable responses are given as a result of inconsistency in interpretation. Microsoft Tay chatbot is a popular example of natural language misinterpretation. It was originally released by Microsoft Corporation via Twitter on March 23, 2016 and was shut down after 16 hours of its launch due to inconsistencies in interpretation [131].

3) Detecting and maintaining conversation. The flow of the conversation is determined by the context of the conversation. Open-domain conversation is another challenge that required determining the topic and keeping track of the context besides detecting when the topic is changed. Despite the spectacular development of chatbots, they are occasionally unable to recognize the intent of users, which makes users frustrated.

4) Data efficiency and time. There is still a problem of little or no datasets being available in certain domains, especially clear in task-oriented systems, where gathering datasets would be costly and time-consuming for new domains. In English, several lexicons are built, such as SentiWordNet and WordNet [132]. By contrast, a lack of Arabic lexicons and resources affects the progress of Arabic chatbots.

5) Evaluation. No universal framework is to evaluate chatbots. There is a lack of unified definitions, metrics, and validated scales in evaluation [133]. The lack of a common

framework makes accurate testing and comparison of different models difficult. Although human evaluation offers qualitative estimation, it is subjective and time-consuming.

6) Privacy and security. Some chatbots are designed to rely on APIs to obtain information, which makes it important to secure the user's data. The kinds of data that the chatbot collects and provides can be used to supplement the support offered by different services, such as in the education domain. This will almost certainly improve the quality and efficiency of the resources available. Moreover, ethical issues of using chatbots are, either from the user side when they abuse the chatbot, or from the chatbot side, when they may save and use the user information for different purposes.

7) Lack of emotions. Some chatbots are developed with predefined conversations that limit their linguistic intelligence and make them mechanical and incapable of communicating in a natural manner. Involving these emotions in chatbots gives them human-like communicative behaviors, recognizing the different meaning possibilities of input and then producing more appropriate responses.

C. Open Research Problems

1) Datasets. Although many word sense disambiguation approaches have been developed [134], [135], they typically increase the computational complexity, which may not be a desirable solution. Moreover, most of the selected articles have built their own dataset, especially for the Arabic research, and others build Arabic corpora [136],[137], [138], whereas others used translation techniques [61],[72]. However, these solutions are limited, and there is still an open-domain problem that needs a lot of attention. In addition, a limitation of Arabic annotated data sets is another problem [136]. Thus, working on providing appropriate data sets and making them available for research can be considered a valuable contribution to chatbot research.

2) Evaluation framework. A new comprehensive framework of evaluation should be provided. Building this framework is not an easy task and may be affected by different factors, such as inputs having multiple semantic meanings and the length of the conversation that may be related to the task itself, for example, the flow of direct questions asked by students differs from entertainment conversations. The framework may also distinguish from different domains, for example, evaluating student understanding differs from completing booking tasks. Thus, the evaluation frameworks must assess providing these tasks to users, and research must define a robust evaluation framework that can mitigate the negative effects of these challenges.

3) Human-like conversation. The current research suffers from limitations in generating a natural conversation, resulting in a noneffective chatbot [10]. A number of factors may affect the behavior of chatbot conversation. First, ambiguous inputs need to be verified and detected to produce appropriate behaviors, such as asking to rephrase the input. Second, bi-linguistic, or multi-linguistic chatbots offer a wider variety of capabilities and provide more user trust. Third, leveraging emotional and contextual cues encourages a user to continue chatting, necessitating further investigation by the researcher into sentiment and emotional analysis. From the Arabic language side, many articles have been conducted on Arabic morphological analysis and generation using a range of methods. This applies in various levels of linguistic complexity, including stemmers [139],[140]. However, Arabic is a derivational and inflectional language that needs a lot of attention and improvement.

4) Extended and different perspectives. Although the selected articles critically survey the state-of-the-art solutions, they focus on specific research. Hence, other perspectives are not addressed in this SLR, and the challenges exposed in this SLR can inspire researchers to focus more on addressing them. Furthermore, due to access constraints, six databases are considered for selecting the articles. Thus, articles in the remaining databases may support or limit the findings in this SLR. Even so, this SLR can be considered as a contribution toward English and Arabic research for the taken criteria.

5) Empirical investigation. This SLR addresses the four research questions without empirical investigation. Involving empirical contributions in future works would give a broader analysis of chatbot development and usage.

V. Conclusion

Chatbot usage has become increasingly prevalent in recent years. One of the key goals of adopting chatbots is minimizing human involvement. The rapid technological advancements of AI aid in achieving this goal and help in the development of more flexible chatbots that are able to produce human-like conversation. This research provides a systematic review of the articles on the evolution of chatbots to investigate the four research questions. A systematic review protocol was used to analyze 50 and 38 articles for English and Arabic works, respectively, and to extract research from six well-known digital databases in computer science. The SLR analyzes the articles in terms of development techniques, domains, evaluation metrics and underlines some challenges and open problems of chatbot development. Furthermore, presenting future directions may assist researchers in identifying crucial aspects that require deep investigation and more development. The findings show that the research domain targeting education receives greater attention from researchers than other domains, and the retrieval-based approach is the most widely utilized approach in this domain. However, this approach is not able to generate a new response that is not predefined in the chatbot's knowledge base. In contrast, the generation-based approach is suitable for tasks that demand providing a new response. Hybrid approaches generally combine between these approaches and are most used for ranking the multiple possible responses when its performance may improve by using one approach rather than another one. However, they still require further developed.

This SLR concludes that current chatbots are still unable to simulate human conversation. Simultaneously, increasing research interest and rapid technological advancements could

evolve chatbot conversation and make chatbots more flexible, fluent, and human-like. Indeed, this SLR provides various recommendations for future articles, which creates a chance for researchers to continue to develop research on chatbots.

ACKNOWLEDGMENT

Lubna Alhenaki would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2022-258.

REFERENCES

[1] M. McTear, "Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots," Synthesis Lectures on Human Language Technologies, vol. 13, no. 3, pp. 1–251, Oct. 2020, doi: 10.2200/S01060ED1V01Y202010HLT048.

[2] K. Darwish et al., "A Panoramic Survey of Natural Language Processing in the Arab World," arXiv:2011.12631 [cs], Nov. 2020, Accessed: Dec. 15, 2020. [Online]. Available: http://arxiv.org/abs/2011.12631

[3] E. H. Almansor and F. K. Hussain, "Survey on Intelligent Chatbots: State-of-the-Art and Future Research Directions," in Complex, Intelligent, and Software Intensive Systems, vol. 993, L. Barolli, F. K. Hussain, and M. Ikeda, Eds. Cham: Springer International Publishing, 2020, pp. 534–543. doi: 10.1007/978-3-030-22354-0_47.

[4] R. Kumar and M. M. Ali, "A Review on Chatbot Design and Implementation Techniques," vol. 07, no. 02, p. 11, 2020.

[5] E. Adamopoulou and L. Moussiades, "An Overview of Chatbot Technology," in Artificial Intelligence Applications and Innovations, vol. 584, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds. Cham: Springer International Publishing, 2020, pp. 373–383. doi: 10.1007/978-3-030-49186-4_31.

[6] M. W. Ashfaque, S. Tharewal, S. Iqhbal, and C. N. Kayte, "A Review on Techniques, Characteristics and approaches of an intelligent tutoring Chatbot system," in 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), 2020, pp. 258–262.

[7] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," Computers and Education: Artificial Intelligence, vol. 2, p. 100033, 2021, doi: 10.1016/j.caeai.2021.100033.

[8] A. M., K. Ramasamy, S. G., and K. S.R., "A Systematic Survey of Cognitive Chatbots in Personalized Learning Framework," in 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, Mar. 2021, pp. 241–245. doi: 10.1109/WiSPNET51692.2021.9419403.

[9] S. Mohamad Suhaili, N. Salim, and M. N. Jambli, "Service chatbots: A systematic review," Expert Systems with Applications, vol. 184, p. 115461, Dec. 2021, doi: 10.1016/j.eswa.2021.115461.

[10] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," Machine Learning with Applications, vol. 2, p. 100006, Dec. 2020, doi: 10.1016/j.mlwa.2020.100006.

[11] G. Caldarini, S. Jaf, and K. McGarry, "A Literature Survey of Recent Advances in Chatbots," Information, vol. 13, no. 1, p. 41, Jan. 2022, doi: 10.3390/info13010041.

[12] S. Singh and H. Beniwal, "A survey on near-human conversational agents," Journal of King Saud University - Computer and Information Sciences, p. S1319157821003001, Nov. 2021, doi: 10.1016/j.jksuci.2021.10.013.

[13] A. Farghaly and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," ACM Transactions on Asian Language Information Processing, vol. 8, no. 4, pp. 1–22, Dec. 2009, doi: 10.1145/1644879.1644881.

[14] M. Hijjawi and Y. Elsheikh, "Arabic Language Challenges in Text Based Conversational Agents Compared to The English Language," IJCSIT, vol. 7, no. 3, pp. 1–13, Jun. 2015, doi: 10.5121/ijcsit.2015.7301.

[15] S. AlHumoud, A. Al, and W. Aldamegh, "Arabic Chatbots: A Survey," ijacsa, vol. 9, no. 8, 2018, doi: 10.14569/IJACSA.2018.090867.

[16] A. A. Elmadany, S. M. Abdou, and M. Gheith, "A Survey of Arabic Dialogues Understanding for Spontaneous Dialogues and Instant Message," IJNLC, vol. 4, no. 2, pp. 75–94, Apr. 2015, doi: 10.5121/ijnlc.2015.4206.

[17] E. S. AL-Hagbani and M. B. Khan, "Support of Existing Chatbot Development Framework for Arabic Language: A Brief Survey," in 5th International Symposium on Data Mining Applications, vol. 753, M. Alenezi and B. Qureshi, Eds. Cham: Springer International Publishing, 2018, pp. 26–35. doi: 10.1007/978-3-319-78753-4_3.

[18] A. Fuad and M. Al-Yahya, "Recent Developments in Arabic Conversational AI: A Literature Review," IEEE Access, vol. 10, pp. 23842–23859, 2022, doi: 10.1109/ACCESS.2022.3155521.

[19] A. Ahmed, N. Ali, M. Alzubaidi, W. Zaghouani, A. Abd-alrazaq, and M. Househ, "Arabic Chatbot Technologies: A Scoping Review," Computer Methods and Programs in Biomedicine Update, p. 100057, Apr. 2022, doi: 10.1016/j.cmpbup.2022.100057.

[20] D. Jurafsky and J. Martin, Speech and Language Processing. 2022.

[21] S. Hussain, O. Ameri Sianaki, and N. Ababneh, "A Survey on Conversational Agents/Chatbots Classification and Design Techniques," in Web, Artificial Intelligence and Network Applications, vol. 927, L. Barolli, M. Takizawa, F. Xhafa, and T. Enokido, Eds. Cham: Springer International Publishing, 2019, pp. 946–956. doi: 10.1007/978-3-030-15035-8_93.

[22] R. Agarwal and M. Wadhwa, "Review of State-of-the-Art Design Techniques for Chatbots," SN COMPUT. SCI., vol. 1, no. 5, p. 246, Sep. 2020, doi: 10.1007/s42979-020-00255-3.

[23] V. V., J. B. Cooper, and R. L. J., "Algorithm Inspection for Chatbot Performance Evaluation," Procedia Computer Science, vol. 171, pp. 2267–2274, 2020, doi: 10.1016/j.procs.2020.04.245.

[24] K. Ramesh, S. Ravishankaran, A. Joshi, and K. Chandrasekaran, "A Survey of Design Techniques for Conversational Agents," in Information, Communication and Computing Technology, vol. 750, S. Kaushik, D. Gupta, L. Kharb, and D. Chahal, Eds. Singapore: Springer Singapore, 2017, pp. 336–350. doi: 10.1007/978-981-10-6544-6_31.

[25] Q. Motger, X. Franch, and J. Marco, "Conversational Agents in Software Engineering: Survey, Taxonomy and Challenges," arXiv:2106.10901 [cs], Jun. 2021, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/2106.10901

[26] I. Sutskever, J. Martens, and G. Hinton, "Generating Text with Recurrent Neural Networks," p. 8.

[27] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," Int. J. Unc. Fuzz. Knowl. Based Syst., vol. 06, no. 02, pp. 107–116, Apr. 1998, doi: 10.1142/S0218488598000094.

[28] "Long Short-Term Memory."

[29] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," arXiv:1412.3555 [cs], Dec. 2014, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/1412.3555.

[30] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," arXiv:1409.1259 [cs, stat], Oct. 2014, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/1409.1259.

[31] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An Empirical Exploration of Recurrent Network Architectures," p. 9.

[32] O. Vinyals and Q. Le, "A Neural Conversational Model," arXiv:1506.05869 [cs], Jul. 2015, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/1506.05869.

[33] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," vol. 2, Jan. 2007.

[34] N. Genc-Nayebi and A. Abran, "A systematic literature review: Opinion mining studies from mobile app store user reviews," Journal of Systems and Software, vol. 125, pp. 207–219, Mar. 2017, doi: 10.1016/j.jss.2016.11.027.

[35] V. Kasinathan, M. H. A. Wahab, S. Z. S. Idrus, A. Mustapha, and K. Z. Yuen, "AIRA Chatbot for Travel: Case Study of AirAsia," J. Phys.: Conf. Ser., vol. 1529, no. 2, p. 022101, Apr. 2020, doi: 10.1088/1742-6596/1529/2/022101.

[36] M. Vanjani, Milam Aiken, and M. Park, "Chatbots for Multilingual Conversations," Jul. 2019, doi: 10.5281/ZENODO.3264011.

[37] "snatchbot." https://snatchbot.me/.

[38] Y. M. Mohialden, M. T. Younis, and N. M. Hussien, "A Novel Approach to Arabic Chabot, Utilizing Google Colab and the Internet of Things: A Case Study at a Computer Center," WEB, vol. 18, no. 2, pp. 946–954, Dec. 2021, doi: 10.14704/WEB/V18I2/WEB18365.

[39] D. Adiwardana et al., "Towards a Human-like Open-Domain Chatbot," arXiv:2001.09977 [cs, stat], Feb. 2020, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/2001.09977.

[40] M. Yang, W. Tu, Q. Qu, Z. Zhao, X. Chen, and J. Zhu, "Personalized response generation by dual-learning based domain adaptation," Neural Networks, vol. 103, pp. 72–82, 2018.

[41] J. Weizenbaum, "ELIZA — A Computer Program For the Study of Natural Language Communication Between Man And Machine," ACM, vol. 9, no. 1, pp. 36–45, 1966.

[42] M. Makatchev et al., "Dialogue patterns of an arabic robot receptionist," in Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction - HRI '10, Osaka, Japan, 2010, p. 167. doi: 10.1145/1734454.1734526.

[43] B. Liu and C. Mei, "Lifelong Knowledge Learning in Rule-based Dialogue Systems," p. 5.

[44] K. Colby, S. Weber, and F. Hilf, "Artificial Paranoia," Artificial Intelligence, pp. 1–25, 1971.

[45] J. Jia, "CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning," Knowledge-Based Systems, vol. 22, no. 4, pp. 249–255, May 2009, doi: 10.1016/j.knosys.2008.09.001.

[46] A. Moubaiddin, O. Shalbak, B. Hammo, and N. Obeid, "Arabic Dialogue System for Hotel Reservation based on Natural Language Processing Techniques," CyS, vol. 19, no. 1, Mar. 2015, doi: 10.13053/cys-19-1-1962.

[47] Navida Belgaumwala, "Chatbot: A Virtual Medical Assistant," International Journal for Research in Applied Science and Engineering Technology, vol. 7, pp. 1042–1050, Jun. 2019, doi: 10.22214/ijraset.2019.6179.

[48] H. Candra, "Designing a Chatbot Application for Student Information Centers on Telegram Messenger Using Fulltext Search Boolean Mode," p. 10.

[49] N. A. I. Omoregbe, I. O. Ndaman, S. Misra, O. O. Abayomi-Alli, and R. Damaševičius, "Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic," Journal of Healthcare Engineering, vol. 2020, pp. 1–14, Sep. 2020, doi: 10.1155/2020/8839524.

[50] O. G. Alobaidi, K. A. Crockett, J. D. O'Shea, and T. M. Jarad, "Abdullah: An Intelligent Arabic Conversational Tutoring System for Modern Islamic Education," p. 7, 2013.

[51] O. G. Alobaidi, K. Crockett, J. D. O'Shea, and T. M. Jarad, "The Application of Learning Theories into Abdullah: An Intelligent Arabic Conversational Agent Tutor:," in Proceedings of the International Conference on Agents and Artificial Intelligence, Lisbon, Portugal, 2015, pp. 361–369. doi: 10.5220/0005197003610369.

[52] S. S. Aljameel, J. D. O'Shea, K. A. Crockett, A. Latham, and M. Kaleem, "Development of an Arabic Conversational Intelligent Tutoring System for Education of children with ASD," in 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Annecy, France, Jun. 2017, pp. 24–29. doi: 10.1109/CIVEMSA.2017.7995296.

[53] S. Aljameel, J. O'Shea, K. Crockett, A. Latham, and M. Kaleem, "LANA-I: An Arabic Conversational Intelligent Tutoring System for Children with ASD," in Intelligent Computing, vol. 997, K. Arai, R. Bhatia, and S. Kapoor, Eds. Cham: Springer International Publishing, 2019, pp. 498–516. doi: 10.1007/978-3-030-22871-2_34.

[54] M. Hijjawi, Z. Bandar, K. Crockett, and D. Mclean, "ArabChat: An arabic conversational agent," in 2014 6th International Conference on Computer Science and Information Technology (CSIT), 2014, pp. 227–237.

[55] M. Hijjawi, Z. Bandar, and K. Crockett, "The Enhanced Arabchat: An Arabic Conversational Agent," ijacsa, vol. 7, no. 2, 2016, doi: 10.14569/IJACSA.2016.070247.

[56] M. Hijjawi, Z. Bandar, and K. Crockett, "User's utterance classification using machine learning for Arabic Conversational Agents," in 2013 5th International Conference on Computer Science and Information Technology, Amman, Jordan, Mar. 2013, pp. 223–232. doi: 10.1109/CSIT.2013.6588784.

[57] M. Hijjawi, Z. Bandar, and K. Crockett, "A Novel Hybrid Rule Mechanism for the Arabic Conversational Agent ArabChat," p. 10, 2015.

[58] M. Hijjawi, H. Qattous, and O. Alsheiksalem, "Mobile Arabchat: An Arabic Mobile-Based Conversational Agent," ijacsa, vol. 6, no. 10, 2015, doi: 10.14569/IJACSA.2015.061016.

[59] Z. Noori, Z. Bandar, and K. Crockett, "Arabic Goal-oriented Conversational Agent Based on Pattern Matching and Knowledge Trees," p. 7, 2014.

[60] Department of Computer Science, College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia, S. M. Yassin, M. Z. Khan, and Department of Computer Science, College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia, "SeerahBot: An Arabic Chatbot About Prophet's Biography," ijircst, vol. 9, no. 2, pp. 89–97, Mar. 2021, doi: 10.21276/ijircst.2021.9.2.13.

[61] N. Mavridis, A. AlDhaheri, L. AlDhaheri, M. Khanii, and N. AlDarmaki, "Transforming IbnSina into an advanced multilingual interactive android robot," in 2011 IEEE GCC Conference and Exhibition (GCC), Dubai, United Arab Emirates, Feb. 2011, pp. 120–123. doi: 10.1109/IEEEGCC.2011.5752467.

[62] N. O. Alshammari and F. D. Alharbi, "Combining a Novel Scoring Approach with Arabic Stemming Techniques for Arabic Chatbots Conversation Engine," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 21, no. 4, pp. 1–21, Jul. 2022, doi: 10.1145/3511215.

[63] S. Z. Sweidan, S. S. Abu Laban, N. A. Alnaimat, and K. A. Darabkh, " SIAAA - C: A student interactive assistant android application with chatbot during COVID - 19 pandemic," Comput Appl Eng Educ, vol. 29, no. 6, pp. 1718‐1742, Nov. 2021, doi: 10.1002/cae.22419.

[64] L. D. Riek et al., "Ibn Sina Steps Out: Exploring Arabic Attitudes Toward Humanoid Robots," p. 8, 2010.

[65] S. Z. Sweidan, S. S. Abu Laban, N. A. Alnaimat, and K. A. Darabkh, "SEG-COVID: A Student Electronic Guide within Covid-19 Pandemic," in 2021 9th International Conference on Information and Education Technology (ICIET), Okayama, Japan, Mar. 2021, pp. 139–144. doi: 10.1109/ICIET51873.2021.9419656.

[66] H. ElGibreen, S. Almazyad, S. B. Shuail, M. A. Qahtani, and L. ALhwiseen, "Robot Framework for Anti-Bullying in Saudi Schools," in 2020 Fourth IEEE International Conference on Robotic Computing (IRC), Taichung, Taiwan, Nov. 2020, pp. 166–171. doi: 10.1109/IRC.2020.00033.

[67] Y. Almurtadha, "LABEEB: Intelligent Conversational Agent Approach to Enhance Course Teaching and Allied Learning Outcomes attainment," JACSM, vol. 13, no. 1, pp. 9–12, 2019, doi: 10.4316/JACSM.201901001.

[68] R. Wallace, "Artificial linguistic internet computer entity (alice)," City, 1995.

[69] Department of Computer Applications Cochin University of Science and Technology Cochin, India and Reshmi. S, "EMPOWERING CHATBOTS WITH BUSINESS INTELLIGENCE BY BIG DATA INTEGRATION," ijarcs, vol. 9, no. 1, pp. 627–631, Feb. 2018, doi: 10.26483/ijarcs.v9i1.5398.

[70] S. Roca, J. Sancho, J. García, and Á. Alesanco, "Microservice chatbot architecture for chronic patient support," Journal of Biomedical Informatics, vol. 102, p. 103305, Feb. 2020, doi: 10.1016/j.jbi.2019.103305.

[71] D. Zhang, X. Chen, Y. Zhang, and S. Qin, "Template-based Chatbot for Agriculture Related FAQs," undefined, 2021, Accessed: May 03, 2022. [Online]. Available: https://www.semanticscholar.org/paper/Template-based-Chatbot-for-Agriculture-Related-FAQs-Zhang-Chen/ec0f4128378064b8014edeb6cbb9cdfa5834ac3a.

[72] B. A. Shawar and E. Atwell, "Accessing an Information System by Chatting," in Natural Language Processing and Information Systems, vol. 3136, F. Meziane and E. Métais, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 407–412. doi: 10.1007/978-3-540-27779-8_39.

[73] B. Shawar and E. Atwell, "An Arabic chatbot giving answers from the Qur'an," in Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles, vol. 2, pp. 197–202, 2004.

[74] B. Abu Shawar, "A Chatbot as a Natural Web Interface to Arabic Web QA," Int. J. Emerg. Technol. Learn., vol. 6, no. 1, pp. 37–43, Mar. 2011, doi: 10.3991/ijet.v6i1.1502.

[75] T. Kadeed, Construction of Arabic Interactive Tool Between Humans and Intelligent Agents. 2014.

[76] D. A. Ali and N. Habash, "Botta: An Arabic Dialect Chatbot," p. 5.

[77] D. Al-Ghadhban and N. Al-Twairesh, "Nabiha: An Arabic Dialect Chatbot," IJACSA, vol. 11, no. 3, 2020, doi: 10.14569/IJACSA.2020.0110357.

[78] N. A. Al-Madi, K. A. Maria, M. A. Al-Madi, M. A. Alia, and E. A. Maria, "An Intelligent Arabic Chatbot System Proposed Framework," in 2021 International Conference on Information Technology (ICIT), Amman, Jordan, Jul. 2021, pp. 592–597. doi: 10.1109/ICIT52682.2021.9491699.

[79] Y. Zhang et al., "Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization," p. 11.

[80] A. M. Bashir, A. Hassan, B. Rosman, D. Duma, and M. Ahmed, "Implementation of A Neural Natural Language Understanding Component for Arabic Dialogue Systems," Procedia Computer Science, vol. 142, pp. 222–229, 2018, doi: 10.1016/j.procs.2018.10.479.

[81] M. Aleedy, H. Shaiba, and M. Bezbradica, "Generating and Analyzing Chatbot Responses using Natural Language Processing," IJACSA, vol. 10, no. 9, 2019, doi: 10.14569/IJACSA.2019.0100910.

[82] O. Octavany and A. Wicaksana, "Cleveree: an artificially intelligent web service for Jacob voice chatbot," TELKOMNIKA, vol. 18, no. 3, p. 1422, Jun. 2020, doi: 10.12928/telkomnika.v18i3.14791.

[83] J. Kapočiūtė-Dzikienė, "A Domain-Specific Generative Chatbot Trained from Little Data," Applied Sciences, vol. 10, no. 7, p. 2221, Mar. 2020, doi: 10.3390/app10072221.

[84] T. Hori, W. Wang, Y. Koji, C. Hori, B. Harsham, and J. R. Hershey, "Adversarial training and decoding strategies for end-to-end neural conversation models," Computer Speech & Language, vol. 54, pp. 122–139, Mar. 2019, doi: 10.1016/j.csl.2018.08.006.

[85] T. Naous, C. Hokayem, and H. Hajj, "Empathy-driven Arabic Conversational Chatbot," p. 11.

[86] K. Palasundram, N. Mohd Sharef, N. A. Nasharuddin, K. A. Kasmiran, and A. Azman, "Sequence to Sequence Model Performance for Education Chatbot," Int. J. Emerg. Technol. Learn., vol. 14, no. 24, p. 56, Dec. 2019, doi: 10.3991/ijet.v14i24.12187.

[87] T. Hu et al., "Touch Your Heart: A Tone-aware Chatbot for Customer Care on Social Media," arXiv:1803.02952 [cs], Mar. 2018, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1803.02952

[88] J. Prassanna, "Towards Building A Neural Conversation Chatbot Through Seq2Seq Model," vol. 9, no. 03, p. 5, 2020.

[89] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing Dialogue Agents: I have a dog, do you have pets too?," arXiv:1801.07243 [cs], Sep. 2018, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1801.07243

[90] R. Patel, N. Bhagora, P. Singh, and K. Namdev, "CLOUD BASED STUDENT INFORMATION CHATBOT," vol. 02, no. 04, p. 4.

[91] T. Alshareef and M. A. Siddiqui, "A seq2seq Neural Network based Conversational Agent for Gulf Arabic Dialect," in 2020 21st International Arab Conference on Information Technology (ACIT), Giza, Egypt, Nov. 2020, pp. 1–7. doi: 10.1109/ACIT50332.2020.9300059.

[92] D. Peng, M. Zhou, C. Liu, and J. Ai, "Human–machine dialogue modelling with the fusion of word- and sentence-level emotions," Knowledge-Based Systems, vol. 192, p. 105319, Mar. 2020, doi: 10.1016/j.knosys.2019.105319.

[93] M. Boussakssou, H. Ezzikouri, and M. Erritali, "Chatbot in Arabic language using seq to seq model," Multimed Tools Appl, vol. 81, no. 2, pp. 2859–2871, Jan. 2022, doi: 10.1007/s11042-021-11709-y.

[94] S. Kim, O.-W. Kwon, and H. Kim, "Knowledge-Grounded Chatbot Based on Dual Wasserstein Generative Adversarial Networks with Effective Attention Mechanisms," Applied Sciences, vol. 10, no. 9, p. 3335, May 2020, doi: 10.3390/app10093335.

[95] V.-K. Tran and L.-M. Nguyen, "Gating mechanism based Natural Language Generation for spoken dialogue systems," Neurocomputing, vol. 325, pp. 48–58, Jan. 2019, doi: 10.1016/j.neucom.2018.09.069.

[96] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805 [cs], May 2019, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805

[97] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," p. 12.

[98] H. Song, Y. Wang, K. Zhang, W.-N. Zhang, and T. Liu, "BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data," arXiv:2106.06169 [cs], Jun. 2021, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/2106.06169

[99] Y. Zhang et al., "DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation," arXiv:1911.00536 [cs], May 2020, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1911.00536.

[100] T. Wael, A. Hesham, M. Youssef, O. Adel, H. Hesham, and M. S. Darweesh, "Intelligent Arabic-Based Healthcare Assistant," in 2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, Oct. 2021, pp. 216–221. doi: 10.1109/NILES53778.2021.9600526.

[101] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot," Computational Linguistics, vol. 46, no. 1, pp. 53–93, Mar. 2020, doi: 10.1162/coli_a_00368.

[102] L. Zhang, Y. Yang, J. Zhou, C. Chen, and L. He, "Retrieval-Polished Response Generation for Chatbot," IEEE Access, vol. 8, pp. 123882–123890, 2020, doi: 10.1109/ACCESS.2020.3004152.

[103] Y. Wu, W. Wu, Z. Li, and M. Zhou, "Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots," arXiv:1805.02333 [cs], May 2018, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/1805.02333

[104] Y. Wu, W. Wu, Z. Li, and M. Zhou, "Response Selection with Topic Clues for Retrieval-based Chatbots," arXiv:1605.00090 [cs], Sep. 2016, Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1605.00090.

[105] Y. S. Wijaya and F. Zoromi, "Chatbot Designing Information Service for New Student Registration Based on AIML and Machine Learning," vol. 1, no. 1, p. 10, 2020.

[106] M. Nuruzzaman and O. K. Hussain, "IntelliBot: A Dialogue-based chatbot for the insurance industry," Knowledge-Based Systems, vol. 196, p. 105810, May 2020, doi: 10.1016/j.knosys.2020.105810.

[107] O. Zahour, "Towards a Chatbot for educational and vocational guidance in Morocco: Chatbot E-Orientation," IJETER, vol. 9, no. 2, pp. 2479–2487, Apr. 2020, doi: 10.30534/ijatcse/2020/237922020.

[108] S. S. Ranavare and R. S. Kamath, "Artificial Intelligence based Chatbot for Placement Activity at College Using DialogFlow," vol. 68, no. 30, p. 10.

[109] S. Sajjapanroj, P. Longpradit, and K. Polanunt, "A Prototype of Google Dialog Flow for School Teachers' Uses in Conducting Classroom Research," p. 14.

[110] W. El Hefny, Y. Mansy, M. Abdallah, and S. Abdennadher, "Jooka: A Bilingual Chatbot for University Admission," in Trends and Applications in Information Systems and Technologies, vol. 1367, Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, and A. M. Ramalho Correia, Eds. Cham: Springer International Publishing, 2021, pp. 671–681. doi: 10.1007/978-3-030-72660-7_64.

[111] Department of Computer Science, Universita Degli Studi di Trento, Italy, A. Fadhil, A. AbuRa'ed, and Information & Communication Technologies, Universitat Pompeu Fabra Barcelona, Spain, "OlloBot - Towards A Text-Based Arabic Health Conversational Agent: Evaluation and Results," in Proceedings - Natural Language Processing in a Deep Learning World, Oct. 2019, pp. 295–303. doi: 10.26615/978-954-452-056-4_034.

[112] S. Memeti and S. Pllana, "PAPA: A parallel programming assistant powered by IBM Watson cognitive computing technology," Journal of Computational Science, vol. 26, pp. 275–284, May 2018, doi: 10.1016/j.jocs.2018.01.001.

[113] R. Alotaibi, A. Ali, H. Alharthi, and R. Almehamdi, "AI Chatbot for Tourist Recommendations: A Case Study in the City of Jeddah, Saudi

Arabia," Int. J. Interact. Mob. Technol., vol. 14, no. 19, p. 18, Nov. 2020, doi: 10.3991/ijim.v14i19.17201.

[114] N. T. M. Trang and M. Shcherbakov, "Enhancing Rasa NLU model for Vietnamese chatbot," vol. 9, p. 6, 2021.

[115] A.-H. Al-Ajmi and N. Al-Twairesh, "Building an Arabic Flight Booking Dialogue System Using a Hybrid Rule-Based and Data Driven Approach," IEEE Access, vol. 9, pp. 7043–7053, 2021, doi: 10.1109/ACCESS.2021.3049732.

[116] "JACOB Voice Chatbot Application using Wit.ai for Providing Information in UMN," IJEAT, vol. 8, no. 6S3, pp. 105–109, Nov. 2019, doi: 10.35940/ijeat.F1017.0986S319.

[117] University of Jeddah, Jeddah, Saudi Arabia and A. A. Qaffas, "Improvement of Chatbots Semantics Using Wit.ai and Word Sequence Kernel: Education Chatbot as a Case Study," IJMECS, vol. 11, no. 3, pp. 16–22, Mar. 2019, doi: 10.5815/ijmecs.2019.03.03.

[118] F. O. Chete and G. O. Daudu, "An Approach towards the Development of a Hybrid Chatbot for Handling Students' Complaints," p. 10.

[119] K. Denecke, S. Vaaheesan, and A. Arulnathan, "A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test," IEEE Trans. Emerg. Topics Comput., vol. 9, no. 3, pp. 1170–1182, Jul. 2021, doi: 10.1109/TETC.2020.2974478.

[120] V. Oguntosin and A. Olomo, "Development of an E-Commerce Chatbot for a University Shopping Mall," Applied Computational Intelligence and Soft Computing, vol. 2021, pp. 1–14, Mar. 2021, doi: 10.1155/2021/6630326.

[121] K. Khavya, "Banking Bot," International Journal of New Technology and Research, vol. 4, no. 7, p. 263023.

[122] K. Mageira, D. Pittou, A. Papasalouros, K. Kotis, P. Zangogianni, and A. Daradoumis, "Educational AI Chatbots for Content and Language Integrated Learning," Applied Sciences, vol. 12, no. 7, p. 3239, Mar. 2022, doi: 10.3390/app12073239.

[123] W.-N. Zhang, Q. Zhu, Y. Wang, Y. Zhao, and T. Liu, "Neural personalized response generation as domain adaptation," World Wide Web, vol. 22, no. 4, pp. 1427–1446, Jul. 2019, doi: 10.1007/s11280-018-0598-6.

[124] A. I. Niculescu, I. Kukanov, and B. Wadhwa, "DigiMo - towards developing an emotional intelligent chatbot in Singapore," in Proceedings of the 2020 Symposium on Emerging Research from Asia and on Asian Contexts and Cultures, Honolulu HI USA, Apr. 2020, pp. 29–32. doi: 10.1145/3391203.3391210.

[125] T. Naous, W. Antoun, R. A. Mahmoud, and H. Hajj, "Empathetic BERT2BERT Conversational Model: Learning Arabic Language Generation with Little Data," arXiv:2103.04353 [cs], Mar. 2021, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/2103.04353.

[126] University Politehnica of Bucharest, A. Grosuleac, S. Budulan, University Politehnica of Bucharest, T. Rebedea, and University Politehnica of Bucharest, "Seeking an Empathy-abled Conversational Agent," in RoCHI - International Conference on Human-Computer Interaction, 2020, pp. 103–107. doi: 10.37789/rochi.2020.1.1.16.

[127] J. Kim, S. Oh, O.-W. Kwon, and H. Kim, "Multi-Turn Chatbot Based on Query-Context Attentions and Dual Wasserstein Generative Adversarial Networks," Applied Sciences, vol. 9, no. 18, p. 3908, Sep. 2019, doi: 10.3390/app9183908.

[128] N. Asghar, I. Kobyzev, J. Hoey, P. Poupart, and M. B. Sheikh, "Generating Emotionally Aligned Responses in Dialogues using Affect Control Theory," arXiv:2003.03645 [cs], Apr. 2020, Accessed: May 01, 2022. [Online]. Available: http://arxiv.org/abs/2003.03645.

[129] "pandorabots." www.pandorabots.com.

[130] M. Canonico and L. D. Russis, "A Comparison and Critique of Natural Language Understanding Tools," CLOUD COMPUTING, p. 7, 2018.

[131] "openai." https://openai.com/api/.

[132] "Learning Multilingual Subjective Language via Cross-lingual Projections," p. 8.

[133] A. B. Kocaballi, L. Laranjo, and E. Coiera, "Understanding and Measuring User Experience in Conversational Interfaces," Interacting with Computers, vol. 31, no. 2, pp. 192–207, Mar. 2019, doi: 10.1093/iwc/iwz015.

[134] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, "A study on similarity and relatedness using distributional and WordNet-based approaches," in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09, Boulder, Colorado, 2009, p. 19. doi: 10.3115/1620754.1620758.

[135] A. Zouaghi, L. Merhbene, and M. Zrigui, "Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation," Artif Intell Rev, vol. 38, no. 4, pp. 257–269, Dec. 2012, doi: 10.1007/s10462-011-9249-3.

[136] C. Lhioui, A. Zouaghi, and M. Zrigui, "A Rule-based Semantic Frame Annotation of Arabic Speech Turns for Automatic Dialogue Analysis," Procedia Computer Science, vol. 117, pp. 46–54, 2017, doi: 10.1016/j.procs.2017.10.093.

[137] C. Lhioui, A. Zouaghi, and M. Zrigui, "The Constitution of an Arabic Touristic Corpus," Procedia Computer Science, vol. 142, pp. 14–25, 2018, doi: 10.1016/j.procs.2018.10.457.

[138] B. A. Shawar and E. S. Atwell, "Using corpora in machine-learning chatbot systems," IJCL, vol. 10, no. 4, pp. 489–516, Nov. 2005, doi: 10.1075/ijcl.10.4.06sha.

[139] K. Taghva, R. Elkhoury, and J. Coombs, "Arabic stemming without a root dictionary," in International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II, Las Vegas, NV, USA, 2005, pp. 152-157 Vol. 1. doi: 10.1109/ITCC.2005.90.

[140] M. Hijjawi, Z. Bandar, K. Crockett, and D. Mclean, "An application of pattern matching stemmer in arabic dialogue system," in KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications, 2011, pp. 35–43.

Duality at Classical Electrodynamics and its Interpretation through Machine Learning Algorithms

Huber Nieto-Chaupis
Universidad Autónoma del Perú
Panamericana Sur Km. 16.3 Villa el Salvador
Lima Perú

***Abstract*—The aim of this paper is to investigate the hypothetical duality of classical electrodynamics and quantum mechanics through the usage of Machine Learning principles. Thus, the Mitchell's criteria are used. Essentially this paper is focused on the radiated energy by a free electron inside an intense laser. The usage of mathematical strategies might be correct to some extent so that one expects that classical equation would contain a dual meaning. The concrete case of Compton scattering is analyzed. While at some quantum field theories might not be scrutinized by computer algorithms, contrary to this Quantum Electrodynamics would constitute a robust example.**

***Keywords*—Classical electrodynamics; quantum mechanics; machine learning principles; mitchell's criteria**

I. Introduction

The science of Machine Learning is been applying to a wide spectrum of disciplines in both basic sciences as well as engineering. Mainly the purpose of this application is the improvement of the functionalities of systems. This is often linked to a kind of optimization of the critic variables of systems. Thus to get the best scenario for each system one needs firstly to identify the relevant pieces that would play a critic role in the chain of processes. Clearly one here would argue that Machine Learning is actually applicable to all those input-output systems whose black-box might be unknown. In this manner emerges the necessity of differentiating the one-way path that do not allows to come back at the beginning of the processes. Among the plethora of Machine Learning philosophies one finds the one invented by Tom Mitchell that establishes that system can learn from a triplet of postulates:

- All system has explicitly a concrete task that allow it to develop in a sustained manner.
- In order to accomplish the nominal task the system must to apply a coherent strategy based on a methodology that would have to exhibit a well-designed performance.
- Once the system has accomplished its task then it would analyze if the performance of the used strategies were the right ones against other alternatives. Only if the task was solved without to expends the system resources then one calculated the efficiency of the involved processes. When this is high enough the one can say that the system has enough learning to be applied successively.

Motivated by the Mitchell's criteria, in this paper the energy radiated by a free electron inside an external super intense laser is treated with these criteria. In essence the study is centered in the following question: Given a relativistic electron in a strong laser, under what conditions the classical physics is abandoned to pass a entire scenario governed by quantum mechanics.

These so-called hybrid theories that combines criteria from classical physics and quantum mechanics have been studied at an entire framework of quantum electrodynamics (QED in short) by R. P. Feynman [1]. Subsequently have appeared the works of Volkov [2], Narozhnyi [3], Vachaspati [4], Kibble [5], Reiss [6], Eberly [7] that have studied QED with infinite waves that can be seen as classical fields without quantization. For example consider the QED Lagrangian:

$$\mathcal{L} = -ie \int dx^4 \bar{\Psi} \gamma_\mu A^\mu \Psi \tag{1}$$

where the spinors $\bar{\Psi}$ and Ψ satisfy the Dirac equation and A^μ the external field expressed as an 4-dimension vector without any type of quantization in contrast with the Dirac spinors. As it is well-known in field theories, from Eq.1 one can extend it to others types of elementary interactions by which is usual to derive the well-known diagrams of Feynman.

This paper is entirely focused on the implications of the vectorial potential $\mathbf{A}$ inside classical electrodynamics that allows to estimate observables such as energy radiated as well as to make predictions at the generation of new sources of powerful light at the super-intense regime. Thus emerges the following questions:

- Under what conditions the energy radiated by an electron is quantized?
- There is an exact boundary that separates the quantum mechanics and classical description of radiation emitted by an relativistic electron?
- Is the interaction electron and external field a system that can be described by the principles of Machine Learning?

This paper explores the capabilities of Machine Learning [8] to measure the limits between classical electrodynamics and QED in the concrete case when a relativistic electron is inside a external super-intense laser. For this end the Mitchell criteria [9] are employed to distinguish the scenarios where a transition from the classical to quantum takes place. In second section the theoretical machinery is presented. In third section the implementation of Mitchell criteria at the classical formalism and its link to QED is presented and discussed. In last section the conclusion of paper is presented.

II. Theoretical Machinery

A. Classical Backscattering Radiation

An important example of the transition from classical to quantum dynamics of light is given by the classical Compton backscattering. Commonly the theory demands to employ the covariant notation given by ϕ=$k^{\mu}x_{\mu}$=$k\cdot x$ based from the definition x_{μ}=(x_0,$\vec{x}$). From this the 4-vector k^{μ}=(1,0,0,1) describes an incident field along the +z direction yielding ϕ=x_0-z. Thus the 4-vector $A_{\mu}\equiv$ (0,$\vec{\mathbf{A}}$) with $\vec{\mathbf{A}}$= $\mathbf{A}_x(\phi)$. For the concrete case of backscattered radiation the direction must be opposite to the incident field and written as -$\hat{k}$, which is means that the emitted radiation travels along the -z direction. With the definition of $\xi=\frac{e^2u_0^2\chi^2}{4\pi^2}$ one can write down:

$$\frac{d^2I(\omega,\mathbf{n}\Rightarrow -z)}{d\omega d\Omega}= \xi\left|\int_{-\infty}^{+\infty}A_x(\phi)\exp\left\{i\chi\left[\phi+\int_{-\infty}^{\phi'}\mathbf{A}^2(\psi)d\psi\right]d\phi\right\}\right|^2. \quad (2)$$

That is the backscattered radiation intensity derived in [10] with χ the shifted Doppler frequency that can be interpreted in classical electrodynamics as the harmonics of radiated energy and at the quantum mechanics language as the number of emitted laser photons. One can see that the quantity $|...|^2$ contains all the information of the processes of classical radiation. Although above nonlinear Compton scattering was derived from Eq. 2, then one can speculate about the possible quantum mechanics that it might contain. Now one can go through the integration of the exponential which is the focus of this paper. In the case of linear polarization the laser field which is assumed to be super-intense is defined as $\mathbf{A}(\psi)$=$a$sin$(\psi)\vec{i}$ then the integrand is written as $a^2\int_{-\infty}^{\phi'}\sin^2(\psi)d\psi$. The integration can be done in a straightforward manner yielding the product of three exponential:

$$\exp\left\{i\chi\left[\phi+\int_{-\infty}^{\phi'}\mathbf{A}^2(\psi)d\psi\right]d\phi\right\} = \exp(i\chi\phi)\exp(i\frac{i\chi a^2}{2}\phi')\exp(-i\frac{\chi a^2}{2}\sin 2\phi') \quad (3)$$

These changes have as objective to create infinite series using the basis of integer-order Bessel's functions guided by the formulation of Ritus and Nishikov [11] as follows:

$$\exp(i\chi\sin\phi)=\sum_{\ell}J_{\ell}(\chi)\exp(i\ell\phi) \quad (4)$$

$$\exp(i\frac{\chi a^2}{2}\sin\phi')=\sum_{m}J_m(\frac{\chi a^2}{2})\exp(im\phi') \quad (5)$$

$$\exp(-i\frac{\chi a^2}{2}\sin 2\phi')=\sum_{n}J_n(\frac{\chi a^2}{2})\exp(-in2\phi') \quad (6)$$

the usage of this technique that favorably ends in a kind of quantization of the intense field but working in a fully QED scenario given by the Volkov's states. Thus, with all these expansions and inserting them in Eq.3 and therefore inserting the result in Eq. 2 then one arrives to an important relation written below as:

$$\frac{d^2I(\omega,-z)}{d\omega d\Omega}=\xi\left|\int_{-\infty}^{+\infty}A_x(\phi)\sum_{\ell}J_{\ell}(\chi)\exp(i\ell\phi)\sum_{m}J_m(\frac{\chi a^2}{2})\exp(im\phi')\sum_{n}J_n(\frac{\chi a^2}{2})\exp(-in2\phi')d\phi\right|^2$$
$$=\xi\left|\int_{-\infty}^{+\infty}A_x(\phi)\sum_{\ell}\sum_{m}\sum_{n}J_{\ell}(\chi)J_m(\frac{\chi a^2}{2})J_n(\frac{\chi a^2}{2})\exp(i\ell\phi)\exp(im\phi')\exp(-in2\phi')d\phi\right|^2. \quad (7)$$

It should be noted that Eq. 7 is a fully classical relation so that in a first instance it is impossible to link it to any fact done inside the quantum mechanics framework. In fact, one can see the it is pure emitted radiation done by a relativistic electron. Nevertheless the work of Ritus [12] it was proposed the connection between a semi-classical description a quantization of external light. In order to follow the Ritus's view the covariant quantities are explicitly written as $\phi'=k'_{\mu}x^{\mu}$ and $\phi=k_{\mu}x^{\mu}$. With this the argument of product of exponentials in Eq. 7 is written as: $i\ell k_{\mu}x^{\mu}+imk'_{\mu}x^{\mu}-2ink'_{\mu}x^{\mu}$. When it is conveniently ordered then one gets $i[\ell k_{\mu}-(2n-m)k'_{\mu}]x^{\mu}$. Clearly one has only information of light either emitted or absorbed despite the fact that the electron is the responsible of these processes. To homogenize the physics of this event it is also convenient to introduce the exponential exp$[i(p_{\mu}-p'_{\mu})x^{\mu}]$. To note that it is possible only if $p_{\mu}-p'_{\mu}\approx 0$ at the space-point x^{μ}. Logically the purpose of this is twofold:

- The conservation of 4-momentum.
- To force a kind of quantization of external field.

In this manner the argument of the resulting product can be written below as:

$$\text{Exp}\left[i(p_{\mu}+\ell k_{\mu}-(2n-m)k'_{\mu}-p'_{\mu}\right)x^{\mu}. \quad (8)$$

Thus one can easily to recognize that there is a pure conservation with the initial and final states of 4-momentum given by:

$$\mathbf{P}_{\text{IN}}=p_{\mu}+\ell k_{\mu}, \quad (9)$$

$$\mathbf{P}_{\text{FI}}=(2n-m)k'_{\mu}+p'_{\mu}. \quad (10)$$

As mentioned above the fact that emerge integer numbers it might not be directly linked to a case of quantization as noted by Ritus. Thus this kind of artificial quantization in a theoretical framework would have to be verified experimentally within a valid window of accuracy [13].

III. The Machine Learning Analysis

Although classical electrodynamics has been entirely developed as a robust branch of physics, the requirement of using advanced algorithms to minimize a biased interpretation of the equations might be seen as an advantage more than a disadvantage at the sense that one gets a kind of hybrid theory with a tuned interpretation. In the following the well-known Mitchell's criteria shall be used to provide a fair interpretation of Eq. 7. These criteria are classified and conceptualized as follows:

- The Task: any system might to have one or more tasks that justifies its existence.
- The Performance: once the task is identified, the system opts by a strategy that must to exhibit a well-defined performance.
- The Experience: depending on the performance and the completion of task, the system should be able to assess the experience along the chain of events. At the cases of an acceptable experience the system can claim that it was a kind of learning to be repeated in subsequent task.

Turning back to Eq. 7, then one can wonder if it requires to isolate a concrete task? And if it is required then for what? One can argue that the manner as Eq. 7 is written, then it represents already a **task** in the sense that one should figure out the best way to extract a reasonable interpretation through a rather clear procedure of integration without any ambiguity. With Eq. 9 and Eq. 10, Eq. 7 can now be written as:

$$\frac{d^2I(\omega,-z)}{d\omega d\Omega}=\xi\left|\int_{-\infty}^{+\infty}A_x(\phi)\sum_{\ell,m,n}J_\ell(\chi)J_m(\frac{\chi a^2}{2})J_n(\frac{\chi a^2}{2})\exp[-i(p_{\rm IN}-p_{\rm FI})_\mu x^\mu]d\phi\right|^2. \tag{11}$$

In addition, the task can also be seen as the demonstration that Eq. 7 is a hybrid formulation of radiated energy by a relativistic free electron. In other words, one can also wonder about the concrete capabilities of classical electrodynamics to exhibit a dual formulation of a fundamental process such as Compton scattering (or it can also be Thomson scattering [14] to some extent, for instance). From Eq. 11 one can wonder if it is quantum mechanics expression or it is needed to corroborate such hypothesis. By following the Mitchell's criteria it is desirable to define a clear route to argument that in fact Eq. 11 is a hybrid expression so that quantum mechanics laws apply. From Eq. 11 to demonstrate that it is a quantum mechanics description of process of interaction between a field and a relativistic free electron, the proposed **performance** requires only the verification at the energy or momentum. For the sake of simplicity the present analysis shall use the energy integration. This demands to employ $\phi = k^\mu x_\mu = \omega t$ that gives $d\phi$=$td\omega$ under the assumption that the photon 4-vector momentum is (1,0,0,0). Thus one arrives to:

$$i[\mathcal{E}+\ell\omega-(2n-m)\omega'-\mathcal{E}']t. \tag{12}$$

With Eq. 12 one clearly finds that its inclusion in Eq.11 turns out to be a Dirac-delta function in the sense that one gets:

$$\int\exp[i(\ell\omega+\mathcal{E}-(2n-m)\omega'-\mathcal{E}')t]\omega dt \\ \omega\delta(\ell\omega+\mathcal{E}-(2n-m)\omega'-\mathcal{E}'). \tag{13}$$

Eq. 13 is an important result in the sense that at least at the energy variable one finds conservation if only if the argument of Dirac-delta function is null. By inserting Eq. 13 into Eq. 11 one arrives to:

$$\frac{d^2I(\omega,-z)}{d\omega d\Omega}=\xi\omega^2a^2\left|\sum_{\ell,m,n}J_\ell(\chi)J_m(\frac{\chi a^2}{2})J_n(\frac{\chi a^2}{2})\delta[(\ell+1)\omega+\mathcal{E}-(2n-m)\omega'-\mathcal{E}']\right|^2. \tag{14}$$

Eq. 14 can be normalized to introduce into a scenario of probabilities. In this manner it is needed that:

$$N^2\frac{d^2I(\omega,-z)}{d\omega d\Omega}=1. \tag{15}$$

Therefore, the constant N can be explicitly written as:

$$N=\sqrt{\frac{1}{\xi\omega^2a^2\left|\sum_{\ell,m,n}J_\ell(\chi)J_m(\frac{\chi a^2}{2})J_n(\frac{\chi a^2}{2})\delta[(\ell+1)\omega+\mathcal{E}-(2n-m)\omega'-\mathcal{E}']\right|^2}}. \tag{16}$$

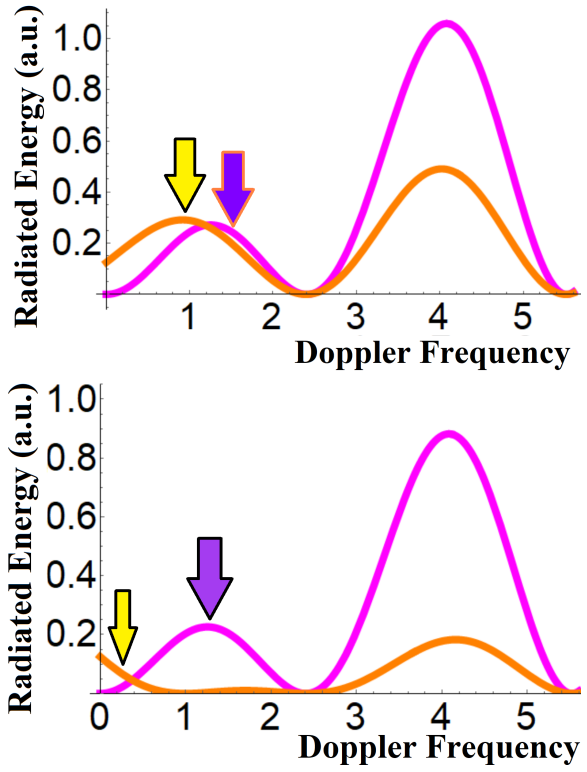


Fig. 1. Classical Distributions of Radiated Energy: Up: $|H(\chi,a)|^2$=0.16$|J_0(\chi)J_1\ (0.001\chi)\ J_0(0.001\chi)|^2$ (Magenta Color), and $|H(\chi,a)|^2$=0.5$|J_0(\chi+0.01)J_1\ (0.01\chi+0.01)\ J_0(0.01\chi+0.01)|^2$ (Orange Color). Down: $|H(\chi,a)|^2$ = 0.52×10^7 $J_0(\chi)\times J_1(0.001\chi)\times J_0(0.001\chi)$ (Orange Color) and $|H(\chi,a)|^2$ = 0.5×10^4 $J_0(\chi-0.01)\times J_1(0.01\chi-0.01)\times J_0(0.01\chi-0.01)$ (Magenta Color).

Once the constant N is calculated then the radiated energy can be estimated fom a procedure based entirely at the Mitchell criteria. Therefore, one can pass the classical electrodynamics concepts to one dictated by principles of Machine Learning.

IV. Interpretation by Mitchell's Criteria

In this way one gets that the applied **performance** has brought elements that play a concrete role at quantum electrodynamics (or QED in short). Thus Eq. 14 makes us to remind the well-known diagrams of Feynman's. In effect, the concordance with QED [15] as seen at the argument of Dirac-delta function has as central implication this pseudo quantization of electromagnetic field that although proposed initially as classically pure, now the interpretation of Feynman's rules [16] would suggest that in the initial state a free electron with initial energy $\mathcal{E}$ absorbs $[\ell+1]$ photons, whereas at the corresponding final state the electron has an energy $\mathcal{E}'$ and has emitted $[2n-m]$ photons. Actually, it is imminently nonlinear Compton scattering as observed by Bula [17]. Since Eq. 14 has been recognized as a potential expression that would play a role in QED, then one expects to arrive to a solid experience after the assumptions that the performance has demanded. In this way one can propose that Eq. 14 is a kind of square of sum of all allowed amplitudes that certainly it could to involve both linear as well as nonlinear contributions [18][19][20][21][22][23][24]. The case of simple Compton is of particular interest. Eq. 14 can be Compton scattering is $\ell=0$ and $2n-m=1$. However one can see that the crude assumption that these integer number would denote the number of photons fails because it is required that $n=m/2$ fact that is totally false in quantum mechanics. This reveals the "**bugs**" of algorithm to propose an effective strategy or performance. Thus simple Compton is restored with $\ell=n=0$ and $m=1$. When $H(\chi,a)$=$J_0(\chi)J_1(\frac{\chi a^2}{2})J_0(\frac{\chi a^2}{2})$ the classical radiated energy can be also interpreted as the measured quantum mechanics observable $\mathcal{O}(\chi,a)$, while $|H(\chi,a)|^2$ the square of all possible amplitudes. Therefore one can write down (where it is assumed after the integration of Dirac-delta function as commonly done in QED):

$$\frac{1}{\xi\omega^2a^2}\frac{d^2I(\omega,-z)}{d\omega d\Omega}=\mathcal{O}(\chi,a)$$
$$=|H_{m=1}(\chi,a)\delta[\omega+\mathcal{E}-\omega'-\mathcal{E}']|^2\approx|H_{m=1}(\chi,a)|^2 \quad (17)$$

A. Generation of Pseudo Amplitudes

In Fig. 1 (Up) one can see the plotting of $|H(1,0,0,\chi,a)|^2$=$|J_0(\chi)J_1(\frac{\chi a^2}{2})J_0(\frac{\chi a^2}{2})|^2$ for two cases: (i) the magenta-color line denoting simple Compton scattering with $\xi\omega^2a^2$=0.1610^7 and $a^2=0.002$ expressing the fact that the laser is not super-intense as initially assumed. (ii) the orange-color line is given by: $|H(1,0,0,\chi,a)|^2$=$|J_0(\chi+0.01)J_1(\frac{\chi a^2}{2}+0.01)J_0(\frac{\chi a^2}{2}+0.01)|^2$ resulting that the peak is shifted to the left-side with a rough coincidence to the value of argument of Bessel function. The added value 0.01 can be perceived as the error at the measurement of Doppler frequency. The reader can corroborate that the case of using -0.01 the negative case, the first peak or Compton peak is gone. In Fig. 1 (Down) the case of $|H(\chi,a)|^2$ = 0.52×10^7 $J_0(\chi)\times J_1(0.001\chi)\times J_0(0.001\chi)$ (orange color) and $|H(\chi,a)|^2$ = 0.5×10^4 $J_0(\chi-0.01)\times J_1(0.01\chi-0.01)\times J_0(0.01\chi-0.01)$ (magenta color) is plotted. While the magenta color distribution exhibits the fact that the central peak is shifted to right-side, the orange color distribution the incorporates errors at the order of 0.01 at the arguments of Bessel function given by: $0.01\chi-0.01$ one can perceive this as the degradation of radiation spectra due to quantum mechanics effects. The product of Bessel functions from above can be defined as a kind of pseudo amplitudes that can be written as:

$$H(m,n,\ell,\chi,a)=\left|J_\ell(\chi)J_m(\frac{\chi a^2}{2})J_n(\frac{\chi a^2}{2})\right|. \quad (18)$$

where $J_\ell(\chi)$ acts as a propagator whereas $J_m(\frac{\chi a^2}{2})$ and $J_n(\frac{\chi a^2}{2})$ can be understood as a kind of input and output states. In fact, the integer number given by the order of Bessel function, is expressing the fact that there is a kind of "classical" absorption as well as emission. From Eq. 17, the square $|H(m,n,\ell,\chi,a)|^2$ represents an observable that is related to radiated energy. This is of importance at the sense that Machine Learning can manage the best values of integer number in order to find the peaks of radiation that to some extent can be perceived as the peaks of X-rays [25][26][27].

In this manner one has below:

$$H(\chi,a)=\sum_{m,n,\ell}\left|J_\ell(\chi)J_m(\frac{\chi a^2}{2})J_n(\frac{\chi a^2}{2})\right|, \tag{19}$$

that emulates to some extent the sum of all possible possibilities for absorption and emission. Then, the intensity of radiated energy is the square of Eq. 19:

$$I(\chi,a)=\left|\sum_{m,n,\ell}J_\ell(\chi)J_m(\frac{\chi a^2}{2})J_n(\frac{\chi a^2}{2})\right|^2. \tag{20}$$

From Eq. 20 one can define the purity of emission at the sense that a detector can sense Compton photons accompanied of pile-up photons, all those that were created by parallel processes. Then this purity is written as:

$$P(\chi,a)=\frac{\left|\sum_{m,n,\ell}J_\ell(\chi)J_m(\frac{\chi a^2}{2})J_n(\frac{\chi a^2}{2})\right|^2}{\left|\sum_{m,n,\ell}J_\ell(\chi)J_m(\frac{\chi a^2}{2})J_n(\frac{\chi a^2}{2})\right|^2+\frac{d^2I_{\rm B}(\omega_{\rm B},a_{\rm B})}{d\omega_{\rm B}d\Omega}} \tag{21}$$

where $I_{\rm B}$, $\omega_{\rm B}$ and $a_{\rm B}$, the intensity, frequency and intensity of background field. In praxis one expects actually that the Compton photons have greater energy than their noise in order to be efficiently detected (see for example [28][29][30][31][32][33]). Thus one has below that:

$$\frac{\frac{d^2I_{\rm B}(\omega_{\rm B},a_{\rm B})}{d\omega_{\rm B}d\Omega}}{\left|\sum_{m,n,\ell}^{M,N,L}J_\ell(\chi)J_m(\frac{\chi a^2}{2})J_n(\frac{\chi a^2}{2})\right|^2}<<1. \tag{22}$$

In this manner, while the Machine Learning **task** would consist in the identification of Compton photons still at a classical scenario. Then, the **performance** would need to have a clear procedure to accomplish the identification of that photons (see for example [34][35][36][37][38]). Thus, it would consist in the searching of the best values of M,N,L to satisfy Eq. 22. Subsequently, the **experience** would be that of Eq. 21. The final Sigmoid function can be derived from Eq. 21 and Eq. 22 so that one arrives to:

$$S(\chi,a,M,N,L)=\frac{1}{1+{\rm Exp}\left[-\frac{\frac{d^2I_{\rm B}(\omega_{\rm B},a_{\rm B})}{d\omega_{\rm B}d\Omega}}{H(\chi,a)}\right]}. \tag{23}$$

The sign "-" in Eq. 23 emerges from the fact that $H(\chi,a)$ can acquire negative values due to the Bessel functions. A deeply analysis of Eq. 23 can be a window to investigate all those available values of classical radiation that might be encompassing quantum mechanics. In addition, the role of these integer number can also be correlated to the existence of a kind of entropy that would appear from the projection of classical observables in a quantum scenario, in the sense that Entropy = Log$[|H(\chi,a)|^{g(r)}]$ (see for example [39], Fig. 2).

V. Conclusion

In this paper, the derivation of quantum observables have been possible with the classical electrodynamics of Hartemann-Kerman equation. The radiated energy equation has been derived, and its relevance in quantum mechanics has been done through the criteria of Tom Mitchell. Along rhis document, integer number have been obtained in a fully analogy to the states of absorption and emission of photons of a relativistic electron in a laser field. Thus, the resulting spectra are strongly dependent on the integer-order Bessel functions. Therefore, the integer number are exploited at the sense that them allow to design a strategy inside the territory of Machine Learning. Finally a Sigmoid function was derived. This clearly demonstrates that classical electrodynamics appears to exhibit a certain flexibility to be adapted to new concepts of computing in physics theoretical as well as experimental.

References

[1] Feynman, Richard P. (1966). Science (August 12, 1966) 153 (3737) 699-708.

[2] D. M. Volkov, Uber eine Klasse von L"osungen der Diracschen Gleichung, Z. Phys. 94, 250 (1935).

[3] NB Narozhnyi, Zh. Eksp. Teor. Fiz., 55: 714-21 (Aug. 1968).

[4] Vachaspati, (1962) Harmonics in the Scattering of Light by Free Electrons. Physical Review, 128 (2). 664.

[5] T. W. B. Kibble, Frequency Shift in High-Intensity Compton Scattering, Phys. Rev. 138, B740, 1965.

[6] Howard R. Reiss and Joseph H: Green's Function in Intense-Field Electrodynamics, Eberly, Phys. Rev. 151, 1058, 1966.

[7] Joseph H. Eberly, Proposed Experiment for Observation of Nonlinear Compton Wavelength Shift, Phys. Rev. Lett. 15, 91, 1965.

[8] Tom Mitchell, Machine Learning , T.M. Mitchell, McGraw Hill, 1997.

[9] T.M. Mitchell, Version Spaces: An Approach to Concept Learning, Ph.D. disseration , Electrical Engineering Department, Stanford University, December, 1978.

[10] F.V. Hartemann and A.K. Kerman, PRL 76, 624 (1996).

[11] A. I. Nikishov and V. I. Ritus, Sovietic Physics, JETP, Vol 2, (19), August 1964.

[12] V. I. Ritus, J. Sov. Laser Res. 6, 497 (1985).

[13] J.D. Franson, Phys. Rev. A 104, 063702 (2021).

[14] G. Fiocco and E. Thompson, Phys. Rev. Lett. 10, 89 (1963).

[15] Huber Nieto-Chaupis, Classical Nonlinear Compton Scattering with Strong Bessel Laser Beams, 2021 IEEE Pulsed Power Conference (2021).

[16] R. P. Feynman: Space-time approach to non-relativistic quantum mechanics. Reviews of Modern Physics 20 (2): 367-387 (1948).

[17] C. Bula and *et.al*, Observation of Nonlinear Effects in Compton Scattering, Phys. Rev. Lett. 76, 3116, (1996).

[18] F. T. Brandt and J. Frenkel, Nonlinear couplings and tree amplitudes in gauge theories, Phys. Rev. D 53, 911 (1996) - Published 15 January 1996.

[19] C. A. Escobar and A. Martín-Ruiz, Equivalence between bumblebee models and electrodynamics in a nonlinear gauge, Phys. Rev. D 95, 095006 (2017) - Published 11 May 2017.

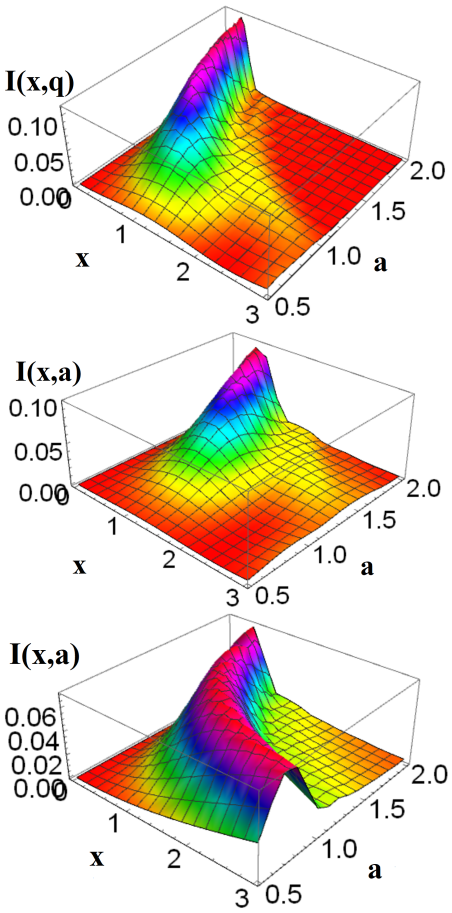


Fig. 2. Illustration of Intensity $I(\chi, a)$=$H(m, n, \ell)$ for $M = N = L$ =2 (Up), $M = N = L = 10$ (Middle) and M, N, L=15 (Down). It is Noteworthy the Formation of a Central Peak Starting from the Lowest Values of Both a as well as the Normalized Doppler-Shifted Frequency.

[20] Jiri Novotny, Self-duality, helicity conservation, and normal ordering in nonlinear QED, Phys. Rev. D 98, 085015 (2018) - Published 17 October 2018.

[21] L. S. Celenza, A. Pantziris, and C. M. Shakin, Chiral symmetry and the nucleon-nucleon interaction: Tensor decomposition of Feynman diagrams, Phys. Rev. C 46, 2213 (1992) - Published 1 December 1992.

[22] D. Bettinelli, R. Ferrari, and A. Quadri, Massive Yang-Mills theory based on the nonlinearly realized gauge group, Phys. Rev. D 77, 045021 (2008) - Published 15 February 2008.

[23] Milton D. Slaughter, Electron-laser pulse scattering, Phys. Rev. D 11, 1639 (1975) - Published 15 March 1975.

[24] Jen-Tsung Hsiang, Tai-Hung Wu, and Da-Shin Lee, Stochastic Lorentz forces on a point charge moving near the conducting plate, Phys. Rev. D 77, 105021 (2008) - Published 21 May 2008.

[25] Yanming Che, Clemens Gneiting, and Franco Nori, Estimating the Euclidean quantum propagator with deep generative modeling of Feynman paths. Phys. Rev. B 105, 214205 (2022) - Published 15 June 2022.

[26] Omry Cohen, Or Malka, and Zohar Ringel, Learning curves for overparametrized deep neural networks: A field theory perspective, Phys. Rev. Research 3, 023034 (2021) - Published 9 April 2021.

[27] Bastian Kaspschak and Ulf-G. Meissner, Neural network perturbation theory and its application to the Born series, Phys. Rev. Research 3, 023223 (2021) - Published 21 June 2021.

[28] Matthew R. Carbone, Shinjae Yoo, Mehmet Topsakal, and Deyu Lu, Classification of local chemical environments from x-ray absorption spectra using supervised machine learning, Phys. Rev. Materials 3, 033604 (2019) - Published 13 March 2019.

[29] Haotong Liang, Valentin Stanev, Aaron Gilad Kusne, Yuto Tsukahara, Kaito Ito, Ryota Takahashi, Mikk Lippmaa, and Ichiro Takeuchi, Application of machine learning to reflection high-energy electron diffraction images for automated structural phase mapping, Phys. Rev. Materials 6, 063805 (2022) - Published 29 June 2022.

[30] Matthew R. Carbone, Mehmet Topsakal, Deyu Lu, and Shinjae Yoo, Machine-Learning X-Ray Absorption Spectra to Quantitative Accuracy, Phys. Rev. Lett. 124, 156401 (2020) - Published 16 April 2020.

[31] O. M. Molchanov, K. D. Launey, A. Mercenne, G. H. Sargsyan, T. Dytrych, and J. P. Draayer, Machine learning approach to pattern recognition in nuclear dynamics from the ab initio symmetry-adapted no-core shell model, Phys. Rev. C 105, 034306 (2022) - Published 3 March 2022.

[32] Pascal Marc Vecsei, Kenny Choo, Johan Chang, and Titus Neupert, Neural network based classification of crystal symmetries from x-ray diffraction patterns, Phys. Rev. B 99, 245120 (2019) - Published 11 June 2019.

[33] Ganesh Sivaraman, Leighanne Gallington, Anand Narayanan Krishnamoorthy, Marius Stan, Gabor Csányi, Alvaro Vazquez-Mayagoitia, and Chris J. Benmore, Experimentally Driven Automated Machine-Learned Interatomic Potential for a Refractory Oxide, Phys. Rev. Lett. 126, 156002 (2021) - Published 14 April 2021.

[34] Yiqun Wang, Xiao-Jie Zhang, Fei Xia, Elsa A. Olivetti, Stephen D. Wilson, Ram Seshadri, and James M. Rondinelli, Learning the crystal structure genome for property classification, Phys. Rev. Research 4, 023029 (2022) - Published 11 April 2022.

[35] IrO_2 Surface Complexions Identified through Machine Learning and Surface Investigations, Jakob Timmermann, Florian Kraushofer, Nikolaus Resch, Peigang Li, Yu Wang, Zhiqiang Mao, Michele Riva, Yonghyuk Lee, Carsten Staacke, Michael Schmid, Christoph Scheurer, Gareth S. Parkinson, Ulrike Diebold, and Karsten Reuter, Phys. Rev. Lett. 125, 206101 (2020) - Published 10 November 2020.

[36] Yuki K. Wakabayashi, Masaki Kobayashi, Yukiharu Takeda, Kosuke Takiguchi, Hiroshi Irie, Shin-ichi Fujimori, Takahito Takeda, Ryo Okano, Yoshiharu Krockenberger, Yoshitaka Taniyasu, and Hideki Yamamoto, Single-domain perpendicular magnetization induced by the coherent O-2p-Ru 4d hybridized state in an ultra-high-quality $SrRuO_3$ film, , Phys. Rev. Materials 5, 124403 (2021) - Published 2 December 2021.

[37] A. H. Lumpkin, R. Thurman-Keup, D. Edstrom, P. Prieto, J. Ruan, B. Jacobson, J. Sikora, J. Diaz-Cruz, A. Edelen, and F. Zhou, Submacropulse electron-beam dynamics correlated with higher-order modes in a Tesla-type cryomodule, Phys. Rev. Accel. Beams 25, 064402 (2022) - Published 24 June 2022.

[38] Juefei Wu, Zili Feng, Jinghui Wang, Qun Chen, Chi Ding, Tong Chen, Zhaopeng Guo, Jinsheng Wen, Youguo Shi, Dingyu Xing, and Jian Sun, Ground states of Au_2Pb and pressure-enhanced superconductivity, Phys. Rev. B 100, 060103(R) (2019) - Published 12 August 2019.

[39] Peter Sollich, Learning from minimum entropy queries in a large committee machine, Phys. Rev. E 53, R2060(R) (1996) - Published 1 March 1996.

Approximate TSV-based 3D Stacked Integrated Circuits by Inexact Interconnects

Mahmoud S. Masadeh
Computer Engineering Department
Yarmouk University
Irbid 21163, Jordan

***Abstract*—Three-Dimensional Stacked Integrated Circuit (3D-SICs) based on Through-Silicon Vias (TSVs) provide a high-density integration technology. However, integrating pre-tested dies requires post-bond interconnect testing, which is complex and costly. An imperfect TSV-based interconnect indicates a defective chip that should be rejected. Thus, it increases the yield loss and test cost. On the other hand, approximate computing (AC) is a promising design paradigm suitable for error-resilient applications, e.g., processing sensory-generated data, by judiciously sacrificing output accuracy. AC perform inexact operations and accepts inexact data. Thus, introducing AC into 3D-SICs will significantly ameliorate the efficiency of design approximation. Therefore, this work aims to *increase the yield and reduce the test cost by accepting 3D-SICs with defected interconnects as approximate 3D-SICs*. This work considers 3D-SICs, where the sensor is stacked on logic (CPU) which is stacked on memory (DRAM). Then, use the memory-based interconnect testing (MBIT) approach to detect and diagnose the faulty interconnect. Based on the detected fault location and type, and for a maximum allowed error, some sensory 3D-SICs with defected LSBs interconnects are accepted and used in error-resilient and data-intensive applications. Targeting data lines only, 50% of the defected interconnects, i.e., least significant bits (LSBs), were accepted as approximate. Thus, the proposed work was able to significantly increase the yield. Two applications, i.e., ECG signal compression and detecting of their R peaks, demonstrated the effectiveness of using a sensory device with a faulty data line in its least significant 8-bits. The approximate ECG signals have a compression rate higher than the exact with negligible (around 0.1%) reduced accuracy.**

Keywords*—Approximate computing; Three-Dimensional Integrated Circuit (3D IC); Through-Silicon Via (TSV); testing; approximate communications; approximate interconnect; yield; energy efficiency

I. Introduction

Three-Dimensional Stacked Integrated Circuit (3D-SICs) based on Through Silicon Vias (TSVs) are emerging among industry and research groups. 3D-SIC is a package with a vertical stack of naked dies which are interconnected utilizing Through-Silicon Vias (TSVs) [1]. TSVs are electrical nails that are etched into the back-side of a thinned-down die, which permit that die to be vertically interconnected to another die. TSVs provide short vertical connections with reduced latency, low capacitance, and low inductance compared to wire-bonds. Thus, TSVs allow for more interconnects between dies with high speed and low power dissipation [2].

The feature-size scaling is becoming difficult and expensive. Moreover, the semiconductor industry is continuously demanding more functionality, bandwidth, and performance at smaller sizes, power dissipation, and cost. Thus, TSV-based 3D-SICs are the promising solution for such requirements [3]. 3D-SICs is a continuation of Moore's Law, which is called *more than Moore's law*. This design paradigm delivers various benefits such as reduced power consumption, reduced footprint, high bandwidth communication, low latency between dies, high transistor density per volume unit, and heterogeneous (e.g., logic, memory, radio frequency (RF) circuits, analog circuits, and sensors) integration [4].

The TSV-based 3D-SICs are promising products for various applications, e.g., the Internet of Things (IoT) and Bio-Medical applications [5]. These applications encompass a tremendous number of mobile and sensory devices, which continuously generate a tremendous quantity of data with redundant and noisy parts. Thus, these data can be processed approximately due to their intrinsic error-resiliency. Similarly, the data could be generated approximately.

Approximate Computing (AC) [6] is an emerging computing paradigm, among both industry and academia, that utilizes the intrinsic resiliency property of Recognition, Mining and Synthesis (RMS) applications. AC provides various benefits such as reducing computation speed, power consumption, and storage space, while achieving an acceptable output quality for various error-resilient applications [7]. Numerous approximation techniques, e.g., voltage over-scaling, approximate arithmetic units, approximate memory, and approximate communication, gained significant interest. However, AC is still immature research direction and does not have standards yet.

Similar to 2D ICs, those TSV-based 3D-SICs require manufacturing testing to meet the expected customer quality. The test operation is executed once at the beginning of the field operation of the IC. Thus, assuming the dies are fault-free, a faulty TSV that could represent a data line, address line, or control line, mandates discarding the whole 3D-SIC. Moreover, workload features could change for an operating IC. Thus, dynamic faults such as Electromigration (EM) should be considered during the operational lifetime. Therefore, to increase the yield and void rejecting an IC with a defected interconnect, *this work proposes accepting TSV-based 3D-SICs with defected interconnects and considering it as an approximate 3D-SICs*. Moreover, extra TSVs could be used to replace the defected interconnects that represent the most significant bits (MSB) of the data and address lines. On the other hand, the error that is caused by the least significant bits (LSB) of the data and address lines can be tolerated without TSV replacement. Extra TSVs are not targeted in this work due to their extra overhead.

The goals while manufacturing DRAM chips differ from those of logic chips, where DRAM designers target reduced area and refresh needs while logic designers target high performance with reduced energy. For best performance, DRAM and logic chips are manufactured individually based on different technology before integration. Thus, wide-IO memory-on-logic are realized as stacked-die applications.

This paper considers 3D-SICs, where the sensor is stacked on memory (DRAM) which is stacked on logic (CPU). Then, use the well-known memory-based interconnect testing (MBIT) approach to detect and diagnose the faulty interconnect. Based on fault location and type, and for a maximum application-dependent acceptable error, some defected 3D-SICs are accepted as approximate. Then, used in error-resilient and data-intensive applications, which tremendously increase the yield rate and reduce test cost.

The rest of this paper is arranged as follows. Section II explains near-sensor computing with various forms of integration. Section III demonstrates TSV fabrication steps, their possible defects and faults, as well as their fault models. The most relevant related work is explained in Section IV. Our proposed methodology is highlighted in Section V. In Section VI, as a case study on ECG signal, we evaluate the proposed methodology and then accept a 3D-SIC with an inexact TSV-based data line. Section VII highlights some of the future directions and concludes the paper.

II. Near-Sensor Computing

The number of sensory devices is expected to reach 75 billion by 2025 and 125 billion by 2030 [8]. They generate a huge amount of repetitious and unformed data. Usually, sensing and processing nodes have different functional requirements and varied manufacturing technology. Moreover, for data sensing, a noisy analog domain is utilized while the data is processed digitally on *von Neumann* computing devices. Thus, sensed data should be transferred from the sensing to the processing node. Therefore, various issues related to response time, data storage, data security, communication bandwidth, and energy consumption should be considered.

There are various forms of integration technologies for near-sensor computing including 3D monolithic, planer SoC, 3D heterogeneous, and 2.5D chiplet integration [9]. In a 3D monolithic integration, the system typically combines various functional layers of sensor, memory and processors in a 3D stacked structure via interlayer vias. For a planer SoC integration, the functional units are integrated with a planar wire connection. However, in 3D heterogeneous integration, different functional units are fabricated individually on different wafers. Then, integrated with advanced packaging technologies, such as TSVs, die-to-die, die-to-wafer and wafer-to-wafer interconnects. This work targets TSVs-based interconnects. For 2.5D integration, the chiplets with specific functions are connected through an interposer, which is a compromise between 2D and 3D packaging integration.

The unprecedented explosion of sensory-generated data and its usage in real-time applications mandates adopting a data-centric approach instead of a computing-centric approach. This enables a system with high performance and energy efficiency. Near-Sensor computing is the solution to provide efficient processing of sensory data with minimal data movement or transformation. In near-sensor computing, the operations of data generation, collection, and processing are performed closed to the sensory devices. The *conventional processing* of sensory-generated data includes data sensing, conversion from analog to digital, storing in memory, transmitting data to the processing unit, then data processing. These steps cause high latency and power consumption. However, the processing units in near-sensor computing reside beside sensors and process data at sensor nodes. Thus, the combination of sensing and computing functions reduces data movement. The sensory computing system performs data processing at two different levels of abstraction, i.e., low and high levels, as described next.

Low-Level Near-Sensor Processing: It removes the undesirable noise from the raw sensory-generated data and includes data filtering, noise suppression and feature enhancement, which are local operations. Such processing ameliorates the computational workload and improves the efficiency of high-level processing. It aims to optimize the features of the raw data. Usually, low-level filtering utilizes circuits located between the sensing devices and high-level processing units.

High-Level Near-Sensor Processing: It comprises the cognitive process that enables the identification of the input signals. It includes recognition, classification and localization. The authors of [10] presented a near-sensor CNN accelerator for image recognition where data processing is close to the sensors. With a near-sensor design, the energy consumption and speed of operation are 60X and 30X, more efficient, respectively, compared to related work. In [11], the authors showed that utilizing 3D stacked ICs (rather than 2D) for near-sensor NN accelerators provides high bandwidth, reduced energy consumption, and low latency of data transfer. Thus, this work targets 3D stacked ICs with inexact TSV-based interconnects.

Near-sensor computing is more complicated than near-memory computing because it includes a huge sensory-generated data of various types. Planer integration of sensors and processing units on a limited area reduces the reserved footprint for sensors. Thus, 3D integration, where sensors are mounted on the top layer while processing units are arranged on the bottom layers, will provide complete exposure for high fill factor. The short distance between the sensing and processing units delivers a high communication bandwidth and low latency. Thus, this work focuses on 3D-SICs by TSV.

III. Interconnect Fault Models

For TSV-based interconnects, this section explains the main used terminology, the basic stages of TSV fabrication, their possible defects, faults, and their fault models.

A. Terminology

Here, we explain various keywords that are used in the rest of the paper. A *defect* is an unintended difference between the implemented hardware and the intended design, emerged from the manufacturing process, e.g., open and bridge defects. The probability of defects in ICs grows with reduced feature size. *Failures* are the physical manifestation of the defect. Defects are generally modeled at a higher conception level

by *faults*, e.g., Stuck-at-Zero (SA0) and Stuck-at-One (SA1). Various defects may be represented with the same fault. A collection of faults with identical properties are grouped in a *fault model*, which should accurately reflect the behavior of defects; as they are used for generating and evaluating test patterns [12]. Faults can be detected by applying a series of *test vectors*; the obtained test responses are compared with golden fault-free responses. The fraction of detectable faults which is called *fault coverage* (FC) indicates the quality of the test.

B. TSV Fabrication Steps

The main manufacturing steps for TSVs, which are cylindrical copper nails, are shown in Fig. 1. These main steps are (*1*) etching of TSV holes: It should be vertical and uniform with a high aspect ratio, (*2*) oxidation: deposition of oxide to isolate the etch from the surrounding semiconductor, (*3*) barrier seed: a barrier layer of metals is deposited before filling the etch with copper. It will prevent the diffusion of the metal into the oxide, (*4*) plating: use copper or tungsten for filling which should be void-free, where the operation of filling should produce minimal stress to avoid warpage, and (*5*) chemical mechanical polishing (CMP): remove the extra layer on the top of the filling. Then, the TSV is ready.

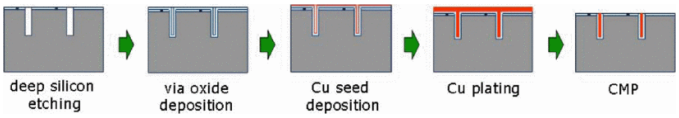


Fig. 1. TSV Fabrication Steps

TSVs can be organized into three classes, based on their fabrication time, during the IC manufacturing process: (*1*) via-first: TSV is fabricated before the front-end logic (FELO, transistor). However, it is more suitable for wafer handing rather than die and it requires adding constraints on design rules of transistor scaling, (*2*) via-middle: TSV is fabricated after the front-end logic (FELO) and before the Back-end of the line (BEOL), which is metal layers deposition, thinning, dicing, and assembly, and (*3*) via-last: TSV is fabricated after the IC fabrication process and before dicing and assembly. It has the lowest TSV fabrication process while being applicable for die and wafer stacking. However, during the manufacturing process, various reasons could cause a defect in the TSV, which are described next.

C. Interconnect Defects, Faults and Fault Models

The various manufacturing steps of TSVs are inherent sources of interconnects defects. There are various defects related to TSV including incomplete fill, pinhole, cracks, TSV misalignment with μ-bumps, TSVs Pinch-off, missing contacts between TSVs and the transistors, and Crosstalk between various TSVs [13].

Fig. 2 shows a general classification of interconnect fault models, which can be static or dynamic. Moreover, a defect can cause a single line or a multi-line fault. SA0 and SA1 are single-line static faults. However, wired-AND and wired-OR are multi-line static faults. Path delay fault (PDF) and path open fault (POF) are single-line dynamic faults. Whenever

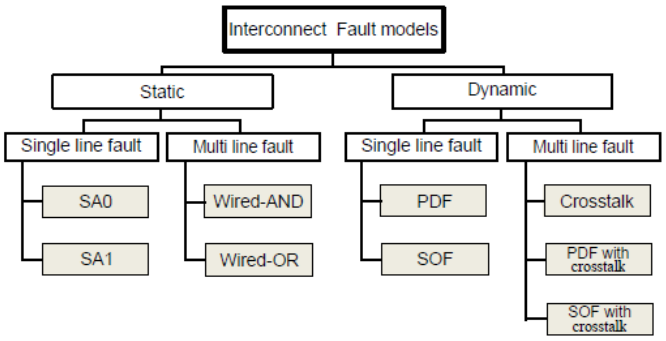


Fig. 2. Classification of Interconnect Fault Models [14]

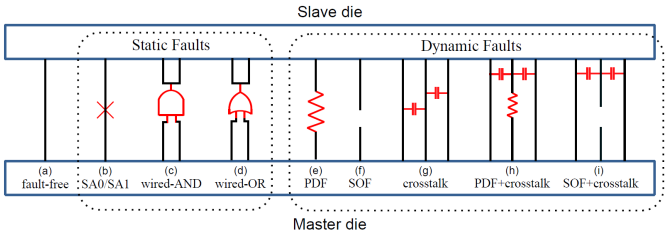


Fig. 3. Static and Dynamic Interconnect Faults [14]

crosstalk is introduced, we will have a multi-line dynamic faults. The interconnect faults are depicted in Fig. 3, including:

1) Stuck-at-Fault (SAF): Has two types which are stuck-at-0 (SA0) and stuck-at-1 (SA1) as depicted in Fig.3(b).

2) Bridge Fault: Simple bridge faults include wired-AND and wired-OR faults. Complex bridge faults also exist, such as (A dominate-AND B) where wire A is fault-free and wire B takes the value A $\cap$ B.

3) Path Delay Fault (PDF): A partial open line defect increases the line delay, e.g., rising or falling delay time (Fig.3(e)).

4) Stuck Open Fault (SOF): This is caused by a completely open line defect (Fig.3(f)).

5) Crosstalk Fault: As shown in Fig.3(g), faults on victim lines are caused by crosstalk from aggressive neighbours. Several crosstalk faults exist as described by the Maximum Aggressor (MA) fault models such as (1) glitch-up, (2) glitch-down, (3) falling delay, and (4) rising delay. Each fault has a specific behavior, while it represents the same phenomena.

6) Path Delay Fault (PDF) with Crosstalk: As shown in Fig.3(h), this is a compound fault where faults due to partial resistive opens are affected by crosstalk from neighbors.

7) Stuck Open Fault (SOF) with Crosstalk: As shown in Fig.3(i), this is a compound fault where faults due to complete open lines are affected by crosstalk from neighbors.

We will show the effects of these faults on TSV-based interconnect and the 3D-SIC as a whole.

IV. Related Work

There is a considerable number of publications that investigate approximate computing, IC testing, and 3D stacked

ICs. However, the portion of the research in approximate computing and hardware design that considers interconnects is scarce. Next, we introduce the most relevant work regarding approximate communication and approximate TSVs.

Recently, researchers investigating various techniques of *approximate communication* for approximate computing. They target network-on-chip (NoCs), aiming for reduced power consumption and latency. The proposed techniques rely on: 1) *lossy compression*: compress each packet and reduce its quality before transmission in order to reduce traffic intensity [15], 2) *value-prediction*: forecast data based on its locality to reduced the transmitted data [16], and 3) *protection-based*: approximate data by protecting the critical part to lower the cost of error correction [17]. These techniques significantly enhance performance and energy consumption. However, controlling the quality of communication is still a significant point. In [18], the authors proposed a hardware-based quality management framework for approximate communication to minimize the time needed for the approximation level calculation. Thus, they presented a new NoC design that observes the application error and adjusts the data approximation level accordingly.

Data transmission across chip interconnects requires a significant amount of time and energy. Thus, the authors of [15] proposed a framework for approximate bus architecture, which is conscious of approximable data. The proposed framework utilizes a light compression technique. For 0.5% quality loss at the application level, the proposed framework achieved a 29% performance improvement. In [19], the authors proposed a framework to reduce power consumption and communication latency of NoCs by incorporating a quality control method and data approximation to reduce packet size. For that, error-resilient variables are identified by analyzing the source code. When transmitting error-resilient variables, a lightweight lossy compression technique is utilized to significantly reduce packet size. In a closely-related work, the same authors explored, in another work, the possibility of using Reinforcement Learning (RL) to manage data quality [20].

The authors of [21], confirmed that the energy consumption of manycore is influenced by data movement, which demands energy-efficient and high-bandwidth interconnects. Towards this direction, they declared that *integrated optics* is an encouraging solution to control the bandwidth limitations of electrical interconnect. However, integrated optics with low-efficiency lasers have high power overhead. Thus, the authors of [21] proposed using *low-power optical signals* to transmit the least significant bits of floating-point numbers. Accordingly, their proposed design has 42% laser power reduction for image processing applications. Similarly, the authors of [22] presented a technique to design scalable *approximate nanophotonic interconnects*. Thus, enhance the interconnect energy efficiency by adjusting the transmission robustness to the application requirements. They achieved a 53% power reduction for output errors of 8%.

The authors of [23] proposed a runtime dynamic Built-In Self-Repair (BISR) technique to improve runtime reliability. For that, they used a test scheme to identify runtime and manufacturing defects. Then, replace defective TSVs with neighbour fault-free TSVs. However, each TSV has its test circuit which causes a large area and power overhead. The authors of [24] showed that testing of 3D-SICs is a challenge due to their complex structure. After stacking, the power and ground TSVs are connected to a grid that makes their testing a challenging task. Thus, they proposed a built-in self-test (BIST) architecture for power and ground TSVs. The proposed BIST enhances reliability by testing for full-open, pin-hole, and bridge faults. However, the proposed BIST introduces hardware overhead with low test coverage.

Previously, various works proposed approximate ICs by designing exact ICs and accepting the defective with minimal fault coverage as approximate ICs [25]. Others proposed designing approximate ICs, and accepting a defective approximate IC if the manufacturing error is within the acceptable approximation error [4]. However, the proposed chips were 2D, not 3D and the approximation is for the logic while considering the interconnect as fault-free. To the authors' knowledge, none of the previous works proposed using ICs with *defective interconnects* as approximate ICs nor targeted designing approximate 3D-SICs, which we propose here. This work mainly targets the communication interconnect itself, i.e., the TSV, as a hardware component. Our proposed idea is a simply different and efficient way. We test and diagnose the faulty TSV-based interconnect with zero area overhead, the ability to detect static and dynamic faults with at-speed testing, and a short test execution time. Then, the output quality of the defected 3D-SIC with defected interconnects is analyzed for a given quality metric. Based on that, some defected 3D-SICs are accepted as approximate ones. Thus, the yield is increased.

V. Proposed Methodology

In this section, we provide a detailed explanation of the proposed methodology. First, we explain Interconnect's built-in self-repair (IBISR). Then, revise memory-based interconnect testing (MBIT). Consequently, the assumed TSV layout is explained because multi-line faults are position-dependent. Next, how a faulty TSV-based data line could be considered approximate is explained.

A. Interconnect Built-In Self-Repair (IBISR)

The researcher of [26] proposed architecture of test and repair of a defect of TSV in 3D-IC, where BIST structure detects a defective TSV. Then, neighbours of the proposed BISR structure isolate and repair the defective TSV. This enhances the yield with an area overhead. The authors of [27] introduced a novel approach for repairing the deficient TSVs in 3D-ICs where interconnect built-in self-test (IBIST) is utilized. Then, the obtained results from IBIST provoke the repairing of defective TSV based on the given BISR structure. They employ repetitious TSV and the time-division multiplexing access (TDMA) in the case of multi defective TSV. However, the high fault rates and TSV footprint make the spare-based repair solutions inadequate [28]. In this work, to keep zero area overhead, we will not introduce interconnect repair. However, it is under investigation for closely related future work.

B. Memory based Interconnect Test (MBIT)

The authors of [14] proposed a Memory Based Interconnect Test (MBIT) approach for 3D-SICs where memory is stacked on logic by testing interconnects through memory read and write operations. MBIT solution can complete at-speed testing

and *diagnosis* and is able to detect all static and dynamic faults. Moreover, MBIT has zero area overhead and allows flexible patterns to be applied. The required test time is much lower than traditional based solutions such as Boundary Scan, but is three times slower than hardwired BIST solutions. However, BIST solutions have a large area overhead and cannot apply flexible patterns. Utilizing MBIT, the minimum set of test patterns required to detect all static and dynamic faults are the patterns to detect *PDF with crosstalk* and *SOP with crosstalk*. We assume a single fault at a time where the number of data lines is L_d = 16, and the number of address lines L_a = 16. We simulate memory test patterns, for a memory die stacked on a logic die that consists of a MIPS64 processor, by using the MIPS64 simulator in [29]. The simulator can handle a maximum of L_d = 64-bit data lines and L_a = 12-bit address lines (lowest 3 bits are byte offset).

C. TSV Layout

The TSV lines represent address, data, and control lines. Also, it include ground and power lines between stacked dies. Testing for multi-line dynamic faults requires knowing the exact layout of the address and data lines. For clarification, we assume a regular TSV array of size 4 × 4 to demonstrate how to generate test patterns for multi-line dynamic faults. Thus, knowing the exact layout is required to accurately analyze the 3D-SIC performance. We assume that a TSV victim is affected by the closest neighbour, i.e., 1^{st} aggressor model. Thus, as shown in Figure 4, a victim TSV (group 1) could be affected with a maximum of 8 aggressors (group 2, 3, and 4).

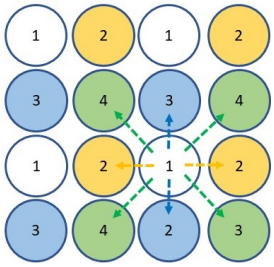


Fig. 4. TSV Layout and Grouping

The JEDEC Solid State Technology Association defends open standards for the microelectronics industry [31]. It provided a standard for stackable Wide-I/O Mobile DRAMs which describes the logic-memory interface for functional and mechanical characteristics, widening the conventional 32-bit DRAM interface to 512 bits. Figure 5 shows the interface for JEDEC Wide-I/O with 1200 connections, where each channel has 300 interconnections. Each channel consists of 6 rows by 50 columns. JEDEC's Wide-I/O interface includes four memory channels, each with 128 bi-directional data lines. Moreover, each channel has 51 control and address signals. Thus, the layout of the interconnections is given where a faulty TSV would be affected by the adjacent ones.

D. The Proposed Methodology

The authors of [32] designed and implemented back-illuminated CMOS image sensors (CIS) (BICIS) with TSV-based bonding between the 3 layers. The number of TSVs for connecting pixel substrate and DRAM substrate is about 15,000 and about 20,000 for connecting the DRAM substrate and the logic substrate. Thus, the fabrication of 35000 TSV could result in defective ones, which will reduce the yield. Therefore, we propose to accept defected TSVs that still provide accepted quality. Algorithm 1 shows the proposed methodology (as a list of steps) for accepting a 3D-SIC with defected TSV-based interconnects as *approximate 3D-SIC*.

Algorithm 1 The Proposed Methodology for Approximate TSV-Based 3D-SICs with Inexact Interconnects

Result: 3D-SIC with inexact interconnects
Manufacture Wafer#1 (CPU);
Manufacture Wafer#2 (DRAM);
Manufacture Wafer#3 (Sensor);
Test and Dice Wafer#1 (CPU);
Test and Dice Wafer#2 (DRAM);
Test and Dice Wafer#3 (Sensor);
Stack the 3 dies by TSVs;
Perform Interconnect Test and Diagnosis;
if *No fault in Address, Data, or Control lines* **then**
    3D-SIC is accepted as Exact; ▷ This is the main goal
else
    if *Control line if faulty* **then**
        3D-SIC is Rejected;
    end if
    if *Data line if faulty* **then**
        Analyze the effect of error based on a given error metric and fault model;
        if *The effect of error is acceptable* **then**
            3D-SIC is Accepted with Faulty Data line; ▷ Data line investigated in this work
        else
            3D-SIC is Rejected;
        end if
    end if
    if *Address line if faulty* **then**
        Analyze the effect of error based on a given error metric and fault model;
        if *The effect of error is acceptable* **then**
            3D-SIC is Accepted with Faulty Address line; ▷ Address line will be investigated in future work
        else
            3D-SIC is Rejected;
        end if
    end if
end if

The wafers of the CPU, DRAM, and Sensor are manufactured then diced. The CPU, DRAM, and Sensor chips are tested at the wafer level and at the die level. Dies stacking is performed through TSV fabrication between the dies. Then, we test the TSV-based interconnects, i.e., MBIT, which implies applying the full list of test patterns. It will detect all possible faults, i.e., various static and dynamic faults. If the obtained test response matches the expected fault-free response for all applied test patterns, the tested 3D-SIC is exact with 100% fault coverage. That represents the ideal case. However, when there is a mismatch between the obtained test response and the expected fault-free response, the tested 3D-SIC is defective, and it should be rejected.

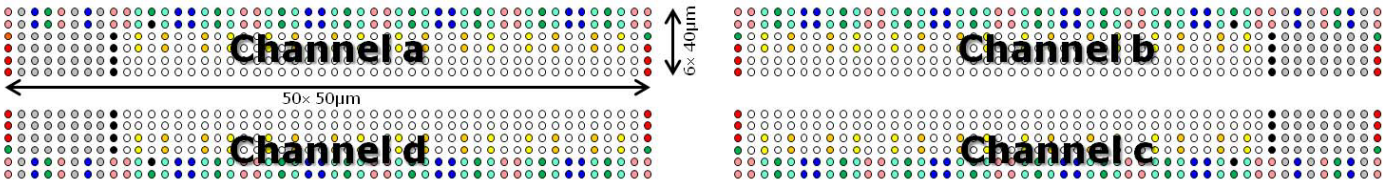


Fig. 5. The Interface for JEDEC Wide-I/O [30]

For a 3D-SIC with defected interconnect, *we perform interconnect diagnosis to identify the exact location and type of the defect*. Rather than discarding the defective 3D-SIC and reducing the yield, we propose to accept some defects based on its location and the used error metric. If a *control line* is identified to be faulty, the operation of the chip will be indeterministic, where it could perform read operation rather than write. Thus, we propose to reject the IC whenever a control line is defective. If a *data* is identified to be faulty, we evaluate its effect on the quality of the final results. If an *address line* is identified to be faulty, this is similar to a faulty address decoder, which is considered as a future work.

In approximate computing, the maximum acceptable error depends on the application, the applied inputs, and user preferences [33]. For that, different error metrics could be used for accuracy evaluation [34] [35], including: (1) Error Rate (ER): which is the percentage of erroneous outputs among all outputs, (2) Error Distance (ED): the arithmetic difference between the exact output and the approximate output for a given input, (3) Mean Error Distance (MED): the average of ED values for a set of outputs obtained by applying a set of inputs, and (4) Relative Error Distance (RED): which is the ratio of ED to the exact output. *Next, we explain how a 3D-SIC with a faulty TSV-based data line could be accepted as approximate IC based on various error metrics*.

E. Faulty Data Line

The number of data lines is L_d = 16 and a faulty data line will be denoted as D_n, for 0 $\leq$ n $\leq$ 15. We assume that the data lines have a normal distribution, where the probability of any line to have a value of 0 or 1 are equal, i.e., $P_{D_n}(0) = P_{D_n}(1)$ = 0.5. Under the assumption that a single fault could occur at a time [12], the error magnitude is 2^n for a faulty D_n data line. The acceptability of a 3D-SIC with a defected interconnect as an approximate one depends on the position of faulty *data line* and the used *accuracy metric*. Next, we explain different error metrics with various fault models:

1: Fault Model is SAF:

Error Metric is ED: For SA0 the data line is always 0, i.e., $P_{D_n}(0)$ = 1, and $P_{D_n}(1)$ = 0. Similarly, for SA1 the data line is always 1, i.e., $P_{D_n}(0)$ = 0, and $P_{D_n}(1)$ = 1. The error magnitude is 2^n for a faulty D_n data line with the assumption of a single fault at a time. Thus, we accept the 3D-SIC as approximate when $2^n >$ ED, and reject it when $2^n \leq$ ED. For large acceptable error, i.e., ED, more chips are accepted as approximate ones. Thus, the yield is increased. When the faulty data line (D_n) is located in the MSB of the design, e.g., 8 $\leq$ n $\leq$ 15, the error magnitude would be large. Thus, the defective chips are rejected, which reduces the yield. On the other hand, when the faulty data line (D_n) is located in the LSB of the design, e.g., 0 $\leq$ n $\leq$ 7, chips are accepted as approximate ones since their error magnitude is small, i.e., $2^7 >$ ED.

Error Metric is ER: The ER indicates the ratio of erroneous outputs among all outputs. A SAF data line, i.e., SA0 or SA1, will give the expected value for 50% of the time and an erroneous result for the rest of the time. Thus, the ER is 50% and the 3D-SIC with defected interconnect is accepted when the allowed ER $\leq$50% and rejected when the ER $>$ 50%.

Error Metric is MED: Under the assumption of one fault at a time the MED metric equals the ED. For multiple errors, MED is given by Equ. 1, where the average ED for a set of faulty data lines is evaluated. The ED for a single data line ranges from $2^{15} = 32K = 32768$ to $2^0 = 1$, based on its location.

$$MED = \frac{1}{16}\sum_{n=0}^{15} ED_n = \frac{1}{16}\sum_{n=0}^{15} 2^n \tag{1}$$

Error Metric is RED: Under the assumption of a single fault at a time the RED for a faulty data line is 1, while it is 0 for an exact interconnect.

2: Fault Model is Bridge Fault (Wired-AND, Wired-OR):

TABLE I. INTERCONNECT VALUE FOR BRIDGE FAULT (WIRED-AND, WIRED-OR)

Exact		A AND B		A OR B	
A	**B**	**A**	**B**	**A**	**B**
0	0	0	0	0	0
0	1	0	**0**	**1**	1
1	0	**0**	0	1	**1**
1	1	1	1	1	1

The bridge fault will give a final value based on: 1) its type; wired-AND or wired-OR, and 2) the value of its neighbour. As shown in Table I, faulty data line with wired-AND will give 0 for 75% of the time and 25% for the rest of the time. Similarly, a faulty data line with wired-OR will give 0 for 25% of the time and give 1 for the rest 75% of the time. Thus, we notice that the bridge fault is mapped to SAF with ER of 25%.

3: Fault Model is Path Delay Fault (PDF):

The dynamic fault of PDF for less than a clock cycle will not cause the circuit to fail. Thus, the data line will deliver an exact value.

4: Fault Model is Stuck Open Fault (SOF):

The SOF represents a completely open line. The floating data line is assumed to have a stable value of 0, a stable value of 1, or changes from 1 to 0. Thus, eventually, the SOF could be equivalent to SAF.

5: Fault Model is Crosstalk:

Figure 4 shows the physical layout of TSVs assuming the 1^{st} aggressor model, where the victim is affected only by the closest neighbour aggressors. Generally, any K^{th} aggressor model can be used, where *K* indicates the maximum TSV distance between victim and aggressors. The authors of [36] showed that restricting *K* to 1 is sufficient.

5.1: PDF with Crosstalk:

A transition at the victim, e.g., from 1 to 0, will be affected by the opposite transition, e.g., from 0 to 1, at the neighbours. Thus, the effect of crosstalk is similar to PDF.

5.2: SOF with Crosstalk:

Detecting SOF with crosstalk requires causing a transition on the victim while keeping the aggressors unchanged. The effect of this model is equivalent to SAF.

F. Possible Repair Scheme

Post-bond interconnect testing for memory stacked on logic requires special consideration since: 1) the stacked dies have different fabrication labs, 2) memory providers are unwilling to incorporate DFT such as JTAG for interconnect testing, and 3) the used DFT can not provide high coverage for dynamic faults. Generally, TSV repair depends on having extra TSVs. However, to avoid extra hardware we will not use extra TSVs nor perform TSV repair.

VI. Case Study

In this section, we evaluate the proposed methodology which accepts a 3D-SIC with an inexact TSV-based data line. Thus, consider it as approximate 3D-SIC, and utilize it in error-resilient applications where reduced accuracy is tolerated.

Biosignal is a human body variable that can be measured and monitored where it provides information on the health status of individuals. Wearable devices sense and process different crucial signs, e.g., electroencephalography (EEG), Electrocardiography (ECG), electrooculogram (EOG), and electromyography (EMG), and send the data to the cloud or to a smartphone. Various biomedical applications accept minor errors or small quality degradation in the values of the biosignal. Electrocardiogram (ECG) is a non-invasive examination that records and shows the electrical activities produced by heart muscle during a cardiac cycle. The ECG test is a standard clinical mechanism for analyzing abnormal heart rhythms and assessing the general condition of a heart. As shown in Figure 6, each ECG cycle consists of 5 waves called PQRST. A complete ECG is recorded using 10 electrodes capturing 12 leads (signals) to get a total picture of the heart. Next, we explain R-peaks detection of Electrocardiography (ECG) signals and its compression assuming the least significant 8-bits are faulty due to inexact TSV-based data line.

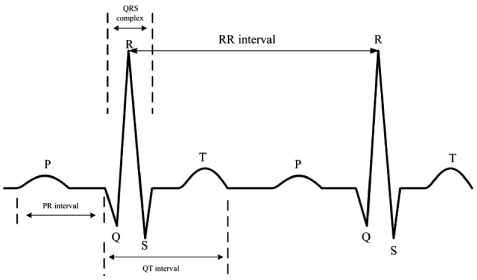


Fig. 6. Parts of an ECG Signal

A. Detecting R-Peaks of ECG Signal

ECG is one of the most critical diagnostic tools for different cardiac diseases. Fast automated detection of the P wave, QRS complex, and T wave is necessary for the early detection of cardiovascular diseases (CVDs). The detection of R-peak is important in all kinds of electrocardiogram (ECG) applications. Utilizing the approach proposed in [37], we performed R peak detection for 32 ECG recordings of the MIT-BIH arrhythmia [38]. For that, we use three parameters, i.e., true-positive (TP), false-negative (FN), and false-positive (FP). TP represents the number of correctly detected R peaks while FN is the number of missed R peaks. FP is the number of noise spikes erroneously classified as R peaks. Utilizing these parameters, we computed various statistical measures including Accuracy (Acc), Precision (positive predictability), Sensitivity/Recall (Se), and F1-Score, as given in the following equations.

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$Precision\ (P) = \frac{TP}{TP + FP} \tag{3}$$

$$Recall/Sensitivity\ (Se) = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

Table II shows the various obtained accuracy metrics, which indicate the high performance of the R peak detection methodology. For the same ECG signals, we created an approximate version of it. For that, the various points of each ECG are approximated by randomly setting one of the least significant 8-bits to zero. This emulates the behaviour of a faulty data line (with SA0 fault) of a sensory device for recording ECG signals.

Stuck-at-0 fault at the least significant data bits did not change the number of total beats, i.e., TP + FN, since the R peak have a high magnitude value. R peak detection of approximate ECG signals missed 57 peaks and classified 123 noise spikes are R Peaks, i.e., FN=57 and FP=123. However, regardless of these false the *accuracy* decreased insignificantly from 99.88% to 99.74%. Similarly, prediction precision, recall, and F1-score reduced insignificantly with less than 0.1%. Thus, a faulty bit in the least significant 8-bits of data line will have

TABLE II. PERFORMANCE OF R PEAK DETECTION METHOD USING ECG EXACT AND APPROXIMATE DATA

	Total beats (TP+FN)	TP (beats)	FN (beats)	FP (beats)	Accuracy	Precision	Recall	F1-Score
Exact ECG	70453	70448	5	77	99.88	99.89	99.99	99.94
Approx ECG	70453	70396	57	123	99.74	99.82	99.91	99.86

reduced effect at application level. Various machine learning-based models could be used as a classifier to detect the QRS complex [39] [40], which are considered as future direction.

B. Compression of Biomedical Signals

Figure 7 shows the architecture for wearable ECG monitoring. The biosignals are acquired, filtered, digitized, *compressed*, and transmitted to the smartphone or cloud server for analysis. The distinct features are obtained, then the classification process detects anomalies. Reducing the amount of transmitted data, through discarding the least significant bits or/and data compression, extends the battery lifetime of mobile devices. Data compression helps to supply the required low-power wireless connection with a slightly large bandwidth.

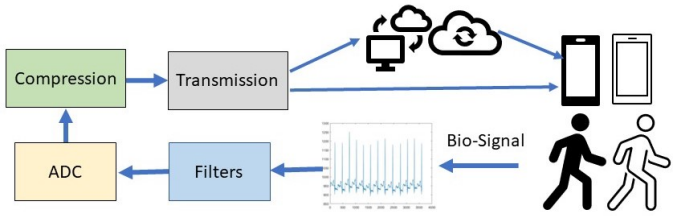


Fig. 7. IoT-Based Wearable ECG Monitoring System

MIT-BIH cardiac arrhythmia database is a widely used database in recent years [38]. MIT-BIH database was supplied by the Massachusetts Institute of Technology with 48 records each is 30 minutes in length. Utilizing the approach proposed in [41], we performed ECG compression for 32 ECG recordings of the MIT-BIH arrhythmia. Then, for the same ECG signals, formed an approximate version of it. Thus, the different points of each ECG are approximated by randomly setting one of the least significant 8-bits to one. This mimics the SA1 fault of a sensory device for recording ECG signals. To assess the ECG signal compression, various metrics are used such as:

Compression Rate (CR): measures the degree of data compression and expressed as given in Eq. 6. Thus, the highest is the best.

Root Mean Squared Error (RMSE): It is a metric for specifying the similarity between two sets, i.e., the original and compressed signal, as expressed in Equ. 7, where *y* is the original signal, *x* is the compressed signal, and *n* is the number of samples of the signal. Thus, the lowest is the best.

$$Compression\ Rate\ (CR) = \frac{Uncompressed\ Size}{Compressed\ Size} \quad (6)$$

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2} \quad (7)$$

TABLE III. RESULTS OBTAINED FOR 32 MIT-BIH RECORDS

	CR	RMSE	Accuracy
Exact ECG	51.7	3.54	99.89
Approx ECG	53.8	6.21	97.91

As shown in Table III, the CR of the exact ECG signal is 51.7 and it is enhanced to 53.8 for the approximate ECG signals. Moreover, the RMSE is 3.54 for the exact ECG signal and it is increased to 6.21 for the approximate ECG which still very acceptable. This work aims to have a high-performance classifier on the compressed signal, both exact and approximate ECG. Thus, the decompressed ECG after lossy compression is classified and detected based on a supporting vector machine (SVM) classifier. The accuracy is 99.89 for the original signal which is reduced insignificantly to 97.91 for the approximate ECG signal. We notice the increase in the compression ratio while keeping the performance of classification of the compressed signal. Thus, a 3D-SIC with a sensory device where the least 8-bits of a data line are faulty can be easily accepted in various applications.

VII. CONCLUSION

Near-sensor computing is a well-known approach to designing efficient hardware for intelligent sensory processing. Data processing at sensory nodes provides a reduced area and time with efficient energy consumption. Thus, it is suitable for real-time and data-intensive applications. However, low-level and high-level near-sensor processing mandates new integration forms and processing algorithms utilizing emerging devices. Although near-senor processing is promising with a great future potential, most of the existing work is still in the development stage and confined to specific applications. This work proposes accepting 3D-SICs with defected TSV-based interconnects as *approximate 3D-SICs*. For this purpose, a sensory device is stacked on a memory die which is stacked on a logic die. To specify if the tested IC is acceptable, context-aware testing is required. Then, a faulty IC is investigated to detect its useability as an approximate one. To evaluate the effectiveness of using a sensory device with a faulty data line in its least significant 8-bits, we performed two applications on ECG signals. First, detecting R peaks of ECG signals then compressing the ECG signals. Both applications demonstrated the usability of the faulty data line in the LSBs of a sensory device. The obtained accuracy metrics, i.e., compression rate, root mean square error, accuracy, precision, recall, and F1-score, showed that a 3D-SIC with a sensory device where the least 8-bits of a data line are faulty can be easily accepted in various applications with enhanced yield.

REFERENCES

[1] P. Garrou, C. Bower, and P. Ramm, *Handbook of 3D integration, volume 1: technology and applications of 3D integrated circuits*. John Wiley & Sons, 2011.

[2] M. Taouil, M. Masadeh, S. Hamdioui, and E. J. Marinissen, "Interconnect test for 3D stacked memory-on-logic," in *Design, Automation Test in Europe Conference Exhibition*, 2014, pp. 1–6.

[3] E. J. Marinissen, *Testing 3D Stacked ICs Containing Through-Silicon Vias*. New York, NY: Springer New York, 2011, pp. 47–74.

[4] M. Masadeh, O. Hasan, and S. Tahar, "Approximation-Conscious IC Testing," in *30th International Conference on Microelectronics*, 2018, pp. 56–59.

[5] U. K. et al., "8Gb 3D DDR3 DRAM using through-silicon-via technology," in *IEEE International Solid-State Circuits Conference - Digest of Technical Papers*, 2009, pp. 130–131,131a.

[6] M. Masadeh, O. Hasan, and S. Tahar, "Comparative Study of Approximate Multipliers," in *Proceedings of the 2018 on Great Lakes Symposium on VLSI*, 2018, pp. 415–418.

[7] ——, "Input-Conscious Approximate Mmultiply-Accumulate (MAC) unit for Energy-Efficiency," *IEEE Access*, vol. 7, pp. 147 129–147 142, 2019.

[8] T. P. Truong, H. T. Le, and T. T. Nguyen, "A reconfigurable hardware platform for low-power wide-area wireless sensor networks," in *Journal of Physics: Conference Series*, vol. 1432, no. 1. IOP Publishing, 2020, p. 012068.

[9] F. Zhou and Y. Chai, "Near-sensor and in-sensor computing," in *Nature Electronics*, vol. 3, 2020, pp. 664–671.

[10] Z. Du, R. Fasthuber, T. Chen, P. Ienne, L. Li, T. Luo, X. Feng, Y. Chen, and O. Temam, "Shidiannao: Shifting vision processing closer to the sensor," in *2015 ACM/IEEE 42nd Annual International Symposium on Computer Architecture (ISCA)*, 2015, pp. 92–104.

[11] T.-H. Hsu, Y.-C. Chiu, W.-C. Wei, Y.-C. Lo, C.-C. Lo, R.-S. Liu, K.-T. Tang, M.-F. Chang, and C.-C. Hsieh, "Ai edge devices using computing-in-memory and processing-in-sensor: From system to device," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 22.5.1–22.5.4.

[12] L.-T. Wang, C.-W. Wu, and X. Wen, *VLSI test principles and architectures: design for testability*. Elsevier, 2006.

[13] M. Masadeh, "Interconnect Testing for 3D Stacked Memories," Delft University of Technology, Delft, Netherlands, 2013.

[14] M. Taouil, M. Masadeh, S. Hamdioui, and E. J. Marinissen, "Post-bond interconnect test and diagnosis for 3-d memory stacked on logic," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 11, pp. 1860–1872, 2015.

[15] J. R. Stevens, A. Ranjan, and A. Raghunathan, "Axba: An approximate bus architecture framework," in *Proceedings of the International Conference on Computer-Aided Design*, 2018, pp. 1–8.

[16] A. Perais and A. Seznec, "Bebop: A cost effective predictor infrastructure for superscalar value prediction," in *IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2015, pp. 13–25.

[17] Y. Chen, M. F. Reza, and A. Louri, "Dec-noc: An approximate framework based on dynamic error control with applications to energy-efficient nocs," in *IEEE 36th International Conference on Computer Design (ICCD)*. IEEE, 2018, pp. 480–487.

[18] Y. Chen and A. Louri, "An online quality management framework for approximate communication in network-on-chips," in *Proceedings of the ACM International Conference on SuperComputing*, 2019, pp. 217–226.

[19] ——, "An approximate communication framework for network-on-chips," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 6, pp. 1434–1446, 2020.

[20] ——, "Learning-based quality management for approximate communication in network-on-chips," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 11, pp. 3724–3735, 2020.

[21] J. Lee, C. Killian, S. L. Beux, and D. Chillet, "Approximate nanophotonic interconnects," in *Proceedings of the 13th IEEE/ACM International Symposium on Networks-on-Chip*. ACM, 2019.

[22] J. Lee, C. Killian, S. Le Beux, and D. Chillet, "Distance-Aware Approximate Nanophotonic Interconnect," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 27, no. 2, nov 2021.

[23] D. K. Maity, S. K. Roy, and C. Giri, "Built-In Self-Repair for Manufacturing and Runtime TSV Defects in 3D ICs," in *IEEE International Test Conference India*, 2020, pp. 1–6.

[24] D. Han, Y. Lee, S. Lee, and S. Kang, "Hardware Efficient Built-in Self-test Architecture for Power and Ground TSVs in 3D IC," in *International SoC Design Conference (ISOCC)*. IEEE, 2021, pp. 101–102.

[25] Z. Jiang and S. K. Gupta, "An ATPG for threshold testing: Obtaining acceptable yield in future processes," in *International Test Conference*. IEEE, 2002, pp. 824–833.

[26] M. Benabdeladhim, A. Fradi, and B. Hamdi, "Interconnect BIST based new self-repairing of TSV defect in 3D-IC," in *International Conference on Engineering MIS (ICEMIS)*, 2017, pp. 1–4.

[27] M. Benabdeladhim, W. DGHAIS, F. ZAYER, and B. Hamdi, "An Efficient Fault Tolerance Technique for Through-Silicon-Vias in 3-D ICs," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 7, pp. 264–270, 2018.

[28] M. Nicolaidis, V. Pasca, and L. Anghel, "I-BIRAS: Interconnect Built-In Self-Repair and Adaptive Serialization in 3D Integrated Systems," in *16th IEEE European Test Symposium*, 2011, pp. 208–208.

[29] "WinMIPS64," last accessed February 7, 2022. [Online]. Available: http://indigo.ie/ mscott/

[30] S. Deutsch, B. Keller, V. Chickermane, S. Mukherjee, N. Sood, S. K. Goel, J.-J. Chen, A. Mehta, F. Lee, and E. J. Marinissen, "DfT architecture and ATPG for Interconnect tests of JEDEC Wide-I/O memory-on-logic die stacks," in *IEEE International Test Conference*. IEEE, 2012, pp. 1–10.

[31] "Wide I/O Single Data Rate (JEDEC Standard JESD229). JEDEC Solid State Technology Association." last accessed March 10, 2022. [Online]. Available: http://www.jedec.org.

[32] Y. Kagawa and H. Iwamoto, "3D Integration Technologies for the Stacked CMOS Image Sensors," in *International 3D Systems Integration Conference (3DIC)*, 2019, pp. 1–4.

[33] M. A. Laurenzano, P. Hill, M. Samadi, S. Mahlke, J. Mars, and L. Tang, "Input Responsiveness: Using Canary Inputs to Dynamically Steer Approximation," in *SIGPLAN Conference on Programming Language Design and Implementation*. ACM, 2016, p. 161–176.

[34] Z. Vasicek and L. Sekanina, "Evolutionary design of approximate multipliers under different error metrics," in *International Symposium on Design and Diagnostics of Electronic Circuits Systems*, 2014, pp. 135–140.

[35] M. Masadeh, O. Hasan, and S. Tahar, "Error analysis of approximate array multipliers," *arXiv preprint arXiv:1908.01343*, 2019.

[36] R. Weerasekera, M. Grange, D. Pamunuwa, H. Tenhunen, and L.-R. Zheng, "Compact modelling of Through-Silicon Vias (TSVs) in three-dimensional (3-D) integrated circuits," in *IEEE International Conference on 3D System Integration*, 2009, pp. 1–8.

[37] J.-S. Park, S.-W. Lee, and U. Park, "R peak detection method using wavelet transform and modified shannon energy envelope," *Journal of healthcare engineering*, vol. 2017, 2017.

[38] "MIT-BIH Arrhythmia Database Directory," last accessed March 25, 2022. [Online]. Available: https://archive.physionet.org/physiobank/database/html/mitdbdir/mitdbdir.htm

[39] M. Masadeh, O. Hasan, and S. Tahar, "Machine learning-based self-compensating approximate computing," in *IEEE International Systems Conference (SysCon)*, 2020, pp. 1–6.

[40] M. Masadeh, Y. Elderhalli, O. Hasan, and S. Tahar, "A Quality-assured Approximate Hardware Accelerators–based on Machine Learning and Dynamic Partial Reconfiguration," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 17, no. 4, pp. 1–19, 2021.

[41] L. Zheng, Z. Wang, J. Liang, S. Luo, and S. Tian, "Effective compression and classification of ecg arrhythmia by singular value decomposition," *Biomedical Engineering Advances*, vol. 2, p. 100013, 2021.

IoT based Low-cost Posture and Bluetooth Controlled Robot for Disabled and Virus Affected People

Tajim Md. Niamat Ullah Akhund[1]
Department of CSE
Daffodil International University
Dhaka, Bangladesh

Mosharof Hossain[2]
Department of CSE
Daffodil International University
Dhaka, Bangladesh

Khadizatul kubra[3]
Department of CSE
Daffodil International University
Dhaka, Bangladesh

Nurjahan[4]
Department of ICT
Bangabandhu Sheikh Mujibur Rahman
Digital University, Bangladesh

Alistair Barros[5]
School of Information Systems
Queensland University of Technology
Brisbane, Australia

Md Whaiduzzaman[6]
School of Information Systems
Queensland University of Technology
Brisbane, Australia

***Abstract*—IoT-based robots can help people to a great extent. This work results in a low-cost posture recognizer robot that can detect posture signs from a disabled or virus-affected person and move accordingly. The robot can take images with the Raspberry Pi camera and process the image to identify the posture with our designed algorithm. In addition, it can also take instructions via Bluetooth from smartphone apps. The robot can move 360 degrees depending on the input posture or Bluetooth. This system can assist disabled people who can move a few organs only. Moreover, this system can assist virus-affected persons as they can instruct the robot without touching it. Finally, the robot can collect data from a distant place and send it to a cloud server without spreading the virus.**

***Keywords*—Internet of Things (IoT); Raspberry Pi; Pi camera; Pi robots; hospital robots; posture recognition robots**

I. Introduction

The Fourth Industrial Revolution (or Industry 4.0), a term coined by Klaus Schwab, refers to the industrial and manufacturing activities automation through using emerging technologies. Robotics as well as the Internet of Things (IoT) is a key element among such modern technologies. In recent times, robots and robotics is the part and parcel of modern science. Robots are working in lots of sectors including the dangerous, monotonous as well as restricted areas where there is a huge risk to human lives. It is a gift of modern artificial Intelligence. People implemented different types of robots for different works. IoT is an ecosystem where things have a unique identifier, embedded with sensors, and actuators, and can send or receive data over the internet. A fusion of both IoT and robotics or in individual manner, these advanced technologies are being used pervasively in many sectors, for example, smart home, agriculture- paddy cultivation [1], smart campus, smart city, environmental weather monitoring, industry automation-poultry farm automation [2], hotel management [3], health areas- patient management [4] as well as taking care of patient by providing nursing facilities [5], remote sensing and monitoring, virus affected and disabled people management [6]. However, the interaction between humans and robots is still a challenging task. Apart from verbal communication, humans use non-verbal signs like postures and gestures to communicate with each other. Non-verbal communication is an instinctive way of communication. It is done by facial expression, gestures/postures, gait, or head nods. It is noteworthy that there is a slight difference between gestures and postures. For example, hand gesture refers to a dynamic state of hand movements whereas hand posture refers to the static state of the hand. In this work, we have proposed an IoT-based system to command a robot with Bluetooth through mobile phone and hand postures, especially beneficial for physically challenged as well as virus-affected people. Apart from this, another system is integrated to measure the body temperature of the person through a contact-less IR temperature sensor and another sensor to measure the temperature and humidity of the room. The collected data will be stored in a secure thingspeak cloud server. Here, hand posture is used as an input to the robot and according to the input, the robot is designed to move towards 360 degrees of movement to left/right/forward/backward and to stop. This robot takes hand posture as input by a pi camera, then converts the posture into commands and performs accordingly.

The objectives of this work are as follows:

1) People can call the robot by hand postures and Bluetooth command to give the robot 360 degrees of movement. As this robot is controlled in two ways, namely, hand posture taken by a pi camera and also mobile Bluetooth, the system is reliable because if one type of controlling fails, the other type may work.
2) This robot may work in virus-affected areas with virus-affected people without making harm to healthy people eventually helping to stop community transmission.
3) The robot can collect data and send them to a cloud server for future monitoring.

The rest of the paper is arranged as follows: Section II discusses the related works associated with this paper, Section III

demonstrates the system model, algorithms, circuit diagrams and flowcharts of the work, Section IV shows the experiment results, comparisons of our work with some other works, limitations and finally Section V provides the concluding remarks and future scope of this work.

II. Literature Review

Researchers explored several approaches to detect hand postures in their works. In literature, the recognition systems are designed based on numerous methods such as deep-learning, 3D Model, depth, skeleton, motion, appearance as well as color [7]. Boyali et al. [8] proposed six different hand gesture and posture (1- Fist, 2- Hand Relax, 3- Fingers Spread, 4- Wave In, 5- Wave Out, 6-Double Tap) recognition system with an accuracy of 97% after receiving electromyography (EMG) signal which is acquired by eight myo armband attached to a person's arm. Here, sparse subspace clustering (SSC) and collaborative representation based classification (CRC) are used to train and recognize patterns. Nguyen et al. [9] demonstrated a hand posture recognition framework consisting of hand detection through traditional hand detector viola jone using haar-like and cascaded adaBoost, low level feature extraction on hand region through image conversion as well as pixel-based extraction, a combination of three kernel descriptors (KDES), namely, gradient, pixel value and texture KDES for hand representation on HSV, RGB, lab color channel and finally multi class support vector machine (SVM) for classification. They have reported an average of 97.3% accuracy on the NUS-2 data-set and 85% on their data-set within 1m to 3m distances. IoT based systems and robots are helping mankind in remote monitoring [10], remote sensing [11], disabled patient management [12], virus affected people management [13] and so on in all times [14] in secure way [15]. Embedded systems and IoT are helping hospitals [16], Agricultural Systems [17], energy generation [18], electronic voting [19] and bio-metric[20] security systems. Akhund et al. [6] reported a 97.9% success rate to recognize hand gestures of disabled as well as virus-affected people to operate the robot. They have developed a robotic agent consisting of an MPU6050 accelero-meter gyroscope sensor, Arduino nano, 433KHz radio wave receiver, L293D motor driver IC. The sensor is responsible for tracking the movement of the hand. Alam et al. [21] designed a hand gesture controlled robot which can recognize five different gestures (left, right, forward, backward, stop) with an accuracy rate of 93.8%. The system uses MPU 6050 module, Arduino nano, HT12E encoder IC, decoder chip HT12D, RF transmitter and receiver. Fakhrurroja et al. [22] developed a system to control electronic devices like fan, light, TV, AC of a smart home through the detection of hand gestures. In their study, kinect v2 sensor is used to track nine different hand state combinations along with raspberry pi and relay module. They have achieved 87% accuracy to detect the gestures. Chen et al. [23] investigated compact CNN named EMGNet to classify hand gestures after analyzing surface electromyography (sEMG) signals obtained through myo armband. The model is tested on Ninapro DB5 Dataset and validate on the myo Dataset. The model can identify 5 different hand gestures with 99.81% recognition accuracy. Maharani et al. [24] implemented two traditional machine learning algorithms namely SVM with directed acyclic graph and K means clustering for the classification of hand postures receiving through Kinect v2. The four gestures (forward, right, left and stop) are used to operate the bioloid premium robot consisting of IC 74LS241N, Arduino mega, kinect v2. They tested hand gesture recognition from the distance of 2, 3 and 4 meters where the range of the body slope was 45, 0 and 45 degrees. They have reported 95.15% accuracy in 10 ms and 77.42% accuracy in 4.45 ms using SVM and k-means clustering algorithm respectively. Meghana et al. [25] designed a system to move a robot towards forward, backward, right and left direction through identifying hand gesture as well as voice especially for the people who are unable to see or hear. The hardware used in the system is MPU6050, L293D driver, LCD, HC05 Bluetooth module, Arduino uno. Abed et al. [26] experimented with a vision-based hand gesture recognition system that can identify five different hand gestures to move a mobile robot towards forward, backward, left, right as well as stop. Raspberry Pi camera module is used to track hand gestures. The hardware that is used to process and provide commands to the robot along with a pi camera is Raspberry Pi 3 model B, L298N motor driver, power supply, camera board, 5 inches 800*480 resistive HD touch screen, rover 5 two-wheel drive platform. They have used Open cv library of python programming language to perform the task. The system showed 98% recognition accuracy. Su designed a 10 types of hand gestures recognition system based on depth vision and surface EMG signals. Here, leap motion controller collects depth vision data finally labeled by hierarchical K means clustering. In addition, myo armband is used to receive and transmit EMG signals. Then, preprocessing is done by a band-stop filter, and a band-pass filter. After feature extraction, finally, multiclass SVM is used to classify the signal. Adithya et al. [27] employed a deep CNN with rectified linear unit (Relu) activation function to automatically recognize hand gestures. They have trained and tested the model using 2 datasets namely the National University of Singapore (NUS) and the American fingerspelling dataset. The model showed 94.7 ± 0.80% accuracy, 94.96 ± 1.20% precision, 94.85 ± 1.30% recall, 94.26 ± 1.70% f1-score in case of NUS dataset and 99.96 ± 0.04% accuracy, 99.96 ± 0.04% precision, 99.96 ± 0.04% recall, 99.96 ± 0.04% f1-Score in case of American fingerspelling dataset. Chansri et al. [28] presented a skin color technique consisting of an RGB camera along with Raspberry Pi to control hand gestures. They have experimented on the American sign language dataset with 12 gestures and achieved 90.83% accuracy. Mondal at el. [29] developed a temperature detector product module for measuring the body temperature of covid 19 affected patients with 98% accuracy. The hardware part consists of LoLin NodeMCU V3 with ESP8266 module, DS18B20 temperature sensor probe, passive buzzer, LED and flat vibrator motor whereas programming is done in Arduino IDE. This system creates an alarm using a buzzer when the temperature exceeds 100.4°F or 38°C temperature. To store the data, they have used ThingSpeak cloud server. IoT based systems are helping people to make cheap irrigation systems [30], Seamless Microservice Execution [31], Big Data Processing [32] and blockchain technology [33]. IoT is helping lung cancer diagnosis [34]. Fog computing plays a vital role for augmenting resource utilization [35]. Cloud computing is helping to analyse people sentiment [36] and dynamic task scheduling algorithms can make it more efficient [37]. So, we should make proper use of IoT and robotics in medical science.

III. Methodology and System Model

A. Requirements

To implement the prototype we used Raspberry Pi 3B+, pi camera 5MP and python 3 programming language for the image processing part. The robot motion controlling part consists of a robotic chassis, breadboard DC gear motor, motor driver L298N, battery, wires, battery charger. Instead, the Arduino UNO Micro-controller, HC-05 Bluetooth Sensor, DHT11 sensor, IR temperature sensor MXL90614, Node MCU ESP8266, Wi-Fi, ThingSpeak cloud server and C++ programming language for Bluetooth controlling and data sending part.

B. System Model

The diagram of the system methodology of the robot movement is depicted in Figure 1. For recognising the posture with image processing we have used Raspberry Pi and Pi camera. The robot can be controlled via Bluetooth also. We could include Bluetooth module with the Raspberry Pi too but we have done it with Arduino uno to make the system more durable and easy. If Raspberry Pi fails any time to detect posture then Arduino will work with following the instructions from mobile.

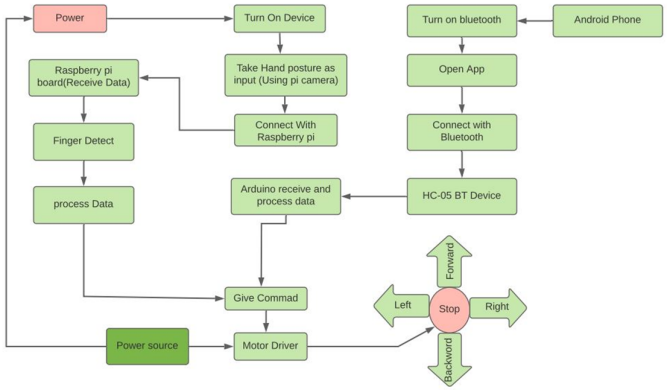


Fig. 1. System Methodology

1) Posture Detection Part: Posture detection is done with image processing using python programming language on Raspberry Pi 3 B+. Pycharm IDE community version was installed on Raspberry Pi. At first, we need to install two library function numpy and opencv. Before plugin library function we updated the pip version 2 to version 3. Then (while (cap.isOpened)) executes video capturing. We crop correct image size. Then we convert color in gray scale and create window for mapping the picture. After that conditions to find motion from the window will be executed with our image processing algorithm. If the motion found then motion pixel detect. We also plot detected motion in window map to simulate the detection in computer screen before applying it in Raspberry Pi and robot. Then the motion data and detected posture creates movement command for the robot to move 360 degrees. The movement commands goes to the GPIO ports of Raspberry Pi and send signal to the motor driver and the motor driver moves the 2 motor of the robot. Finally the robot moves by following the posture signal. If the motion not found then the robot will stop movement and again executes new video reading command. Algorithm for Movement controlling with posture recognition by Raspberry Pi and image processing is as mentioned in Algorithm 1.

Algorithm 1 Posture Recognition with Raspberry Pi and Image Processing

```
import: cv2; numpy ← np; math; cap = cv2.VideoCapture(0)
while (cap.isOpened()) do
    ret, img = cap.read()
    cv2.rectangle(img, (300, 300), (100, 100), (0, 255, 0), 0)
    crop_img = img[100 : 300, 100 : 300]
    grey = cv2.cvtColor(crop_img, cv2.COLOR_BGR2GRAY)
    value = (35, 35)
    blurred = cv2.GaussianBlur(grey, value, 0)
    thresh1 = cv2.threshold(blurred, 127, 255, cv2.T.B.I+ cv2.T.O)
▷ T.B.I=THRESH.BINARY.INV ▷ T.O=THRESH.OTSU
    cv2.imshow('Thresholded', thresh1)
    (version) = cv2.version.split('.')
    if version == '3' then
        image, contours, hierarchy = cv2.findContours()
    else if version == '4' then
        contours, hierarchy = cv2.findContours()
        cnt = max(contours, key=lambda x: cv2.contourArea(x))
        x, y, w, h = cv2.boundingRect(cnt)
        cv2.rectangle(crop.img, (x, y), (x + w, y + h), (0, 0, 255), 0)
        hull = cv2.convexHull(cnt)
        drawing = np.zeros(crop.img.shape, np.uint8)
        cv2.drawContours(drawing, [cnt], 0, (0, 255, 0), 0)
        cv2.drawContours(drawing, [hull], 0, (0, 0, 255), 0)
    end if
    hull = cv2.convexHull(cnt, returnPoints=False)
    defects = cv2.convexityDefects(cnt, hull)
    count.defects = 0
    cv2.drawContours(thresh1, contours, -1, (0, 255, 0), 3)
    for i in range(defects.shape[0]) do
        s, e, f, d = defects[i, 0]
        start = tuple(cnt[s][0])
        end = tuple(cnt[e][0])
        far = tuple(cnt[f][0])
        a = math.sqrt((end[0] - start[0]) ** 2 + (end[1] - start[1]) ** 2)
        b = math.sqrt((far[0] - start[0]) ** 2 + (far[1] - start[1]) ** 2)
        c = math.sqrt((end[0] - far[0]) ** 2 + (end[1] - far[1]) ** 2)
        angle = math.acos((b ** 2 + c ** 2 - a ** 2) / (2 * b * c)) * 57
    end for
    if angle <= 90 then
        count.defects + = 1
        cv2.circle(crop.img, far, 1, [0, 0, 255], -1)
        cv2.line(crop.img, start, end, [0, 255, 0], 2)
    end if
```

```
    if count.defects == 1 then
        str = "Detect 1 Finger"
        cv2.putText(img, str, (50, 50), cv2.F.H.S, 0.5, (0,0, 255), 2)
                    ▷ F.H.S = FONT.HERSHEY.SIMPLEX
    else if count.defects == 2 then
        str = "Detect 2 Finger"
        cv2.putText(img, str, (50, 50), cv2.F.H.S, 0.5, (0,0, 255), 2)
    else if count.defects == 3 then
        str = "Detect 3 Finger"
        cv2.putText(img, str, (50, 50), cv2.F.H.S, 0.5, (0,0, 255), 2)
    else if count.defects == 4 then
        str = "Detect Entire Hand"
        cv2.putText(img, str, (50, 50), cv2.F.H.S, 0.5, (0,0, 255), 2)
    else
        str = "Detect No Finger"
        cv2.putText(img, str, (50, 50), cv2.F.H.S, 0.5, (0,0, 255), 2)
        cv2.imshow('Gesture', img)
        all.img = np.hstack((drawing, crop.img))
        cv2.imshow('Contours', all.img)
        k = cv2.waitKey(10)
    end if
    if k == 27 then
        break
    end if
end while
```

Flow chart of Hand posture recognition and finger detection is mentioned in Figure 2.

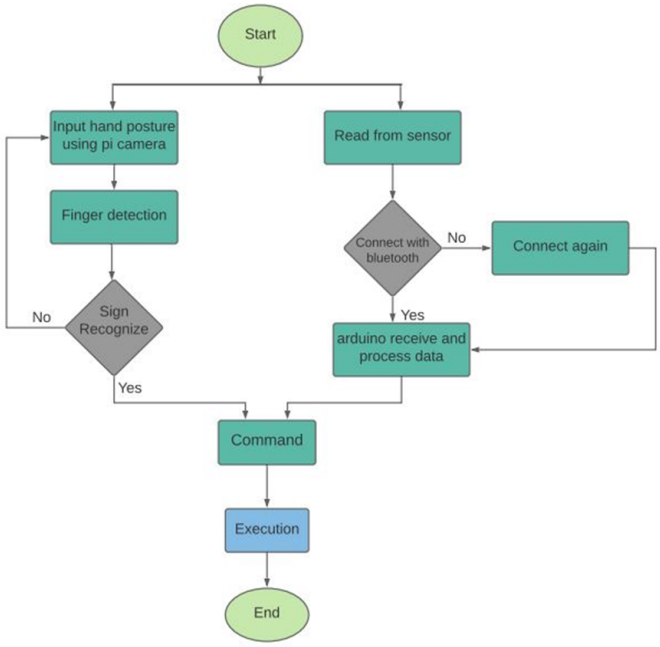


Fig. 2. Flow Chart of Hand Posture Recognition and Finger Detection

The circuit diagram of posture detection and robot controlling with Raspberry Pi is mentioned in Figure 3.

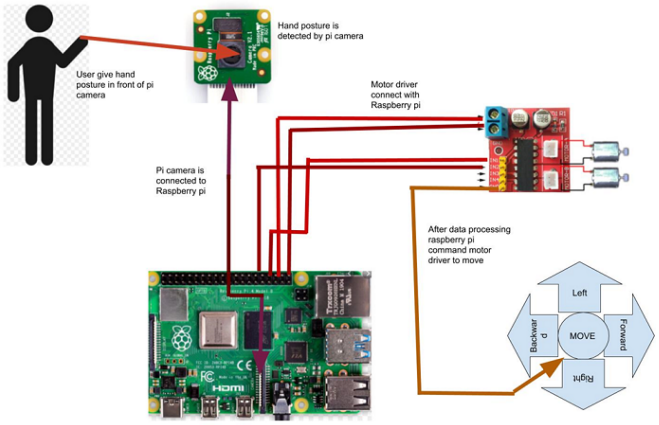


Fig. 3. Circuit Diagram of Posture Detection and Robot Controlling with Raspberry Pi

2) Bluetooth Sensing and Controlling Part: Movement with Bluetooth sensing from mobile phone is another part of this project. We used the Arduino IDE to write the C++ programming language for Arduino UNO Micro-controller. The Arduino will receive the signal sent from mobile phone with Bluetooth sensor HC-05. Android or IOS mobile phones can send any data via any Bluetooth data sender mobile app (we used Bluetooth RC controller available in app store and play store) to HC-05 after pairing. Arduino can receive that data with serial communication (Tx and Rx). After receiving the data in Arduino, conditions for 360 degrees movement will be executed. Then the signals for forward, backward, left, right and stop will be sent to the motor driver L298N. Then the motor driver will move the 2 DC motors of robot by following the commands. Algorithm for Movement Controlling with Bluetooth is mentioned in Algorithm 2.

The circuit diagram of robot controlling with Bluetooth is mentioned in Figure 4.

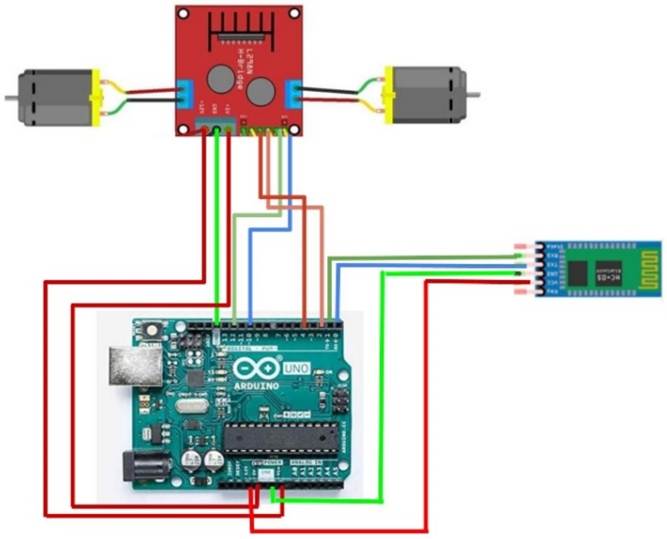


Fig. 4. Circuit Diagram of Robot Controlling with Bluetooth

Algorithm 2 Movement Controlling with Bluetooth

```
Include Libraries.
m11 ← 5; m12 ← 6; m21 ← 9; m22 ← 10; char str[2],i;   ▷ Defining Arduino pins for motor driver and taking variables.
procedure SETUP()                                      ▷ Setup for Arduino
    Serial.begin(9600);
    pinMode(m11, OUTPUT);
    pinMode(m12, OUTPUT);
    pinMode(m21, OUTPUT);
    pinMode(m22, OUTPUT);
end procedure
while Serial.available() do                            ▷ Establishing connection
    char ch=Serial.read();                             ▷ Receiving Bluetooth data
    if (ch=='F') then
        digitalWrite(m11, HIGH);
        digitalWrite(m12, LOW);
        digitalWrite(m21, HIGH);
        digitalWrite(m22, LOW);                        ▷ Going Forward
    else if (ch=='B') then
        digitalWrite(m21, LOW);
        digitalWrite(m22, HIGH);
        digitalWrite(m11, LOW);
        digitalWrite(m12, HIGH);                       ▷ Going Backward
    else if (ch=='L') then
        digitalWrite(m11, HIGH);
        digitalWrite(m12, LOW);
        digitalWrite(m21, LOW);
        digitalWrite(m22, LOW);                        ▷ Going Left
    else if (ch=='R') then
        digitalWrite(m11, LOW);
        digitalWrite(m12, LOW);
        digitalWrite(m21, HIGH);
        digitalWrite(m22, LOW);                        ▷ Going Right
    else if (ch=='S') then
        digitalWrite(m11, LOW);
        digitalWrite(m12, LOW);
        digitalWrite(m21, LOW);
        digitalWrite(m22, LOW);                        ▷ Stop Movement
    end if
end while
```

3) Data Sensing and Sending Part: The system will be able to collect data from the patient with IR temperature sensor MXL90614 without touching people and collect the room temperature and humidity with DHT11 sensor. Then it will upload the collected data to a cloud server of ThingSpeak via Wi-Fi. The collected data will be shown in an Oled display and can be monitored from anywhere in the world with that cloud platform. The circuit concept of data sensing and sending part is shown in Figure 5. We used Node-MCU ESP8266 as micro-controller unit for this.

Algorithm for Data sensing and sending with NodeMCU is mentioned in 3.

IV. Results

We have got experimental result of this hand posture recognition robot. When it detects one finger it goes forward. When the pi camera detects two fingers of hand then it goes

Algorithm 3 Remote Data Sensing with Node MCU

```
Include Libraries: SPI, Wire, Adafruit-GFX, Adafruit-SSD1306, DHT, ESP8266WiFi, WiFiClient and ThingSpeak.
oledRESET ← ledBUILTIN; DHTPIN ← D4; DHTTYPE ← DHT11; ssid ← wifiName; password ← wifiPassword; myChannelNumber ← thingspeakChannelNumber; myWriteAPIKey ← thingspeakWriteApiKey; dataState ← false;   ▷ Defining pins and variables.
AdafruitSSD1306 display(oledRESET);
DHT dht(DHTPIN, DHTTYPE);
procedure SETUP()
    Serial.begin(115200);
    dht.begin();
    if (SSD1306LCDHEIGHT != 64) then
        print("Height incorrect, please fix AdafruitSSD1306.h!");
    end if
    display.begin(SSD1306.SWITCHCAPVCC, 0x3C);
    display.clearDisplay();
    display.display();
    display.setTextSize(1);
    display.setTextColor(WHITE); ▷ making oled display ready
    WiFi.begin(ssid, password);
    while (WiFi.status() != WLCONNECTED) do
        delay(500);
        Serial.print(".");
    end while
    print(WiFi.localIP());      ▷ Connect to WiFi network
    ThingSpeak.begin(client);   ▷ start cloud server communication
end procedure
procedure VOID LOOP()
    temperature = dht.readTemperature();
    humidity = dht.readHumidity();   ▷ getting the data
    display.clearDisplay();
    display.setCursor(0, 0);
    display.print(temperature, " ");
    display.print(humidity);
    display.display();   ▷ printing the data in oled display
    Serial.println(temperature);
    Serial.print(humidity);   ▷ printing the data in serial monitor
    if (temperature < 100 and humidity < 150) then
        if dataState then
            ThingSpeak.writeField(myChannelNumber, 1, temperature, myWriteAPIKey);
            dataState = false;
        else
            ThingSpeak.writeField(myChannelNumber, 2, humidity, myWriteAPIKey);
            dataState = true;
        end if
    end if
    delay(15000);   ▷ uploading the data to the cloud server
end procedure
```

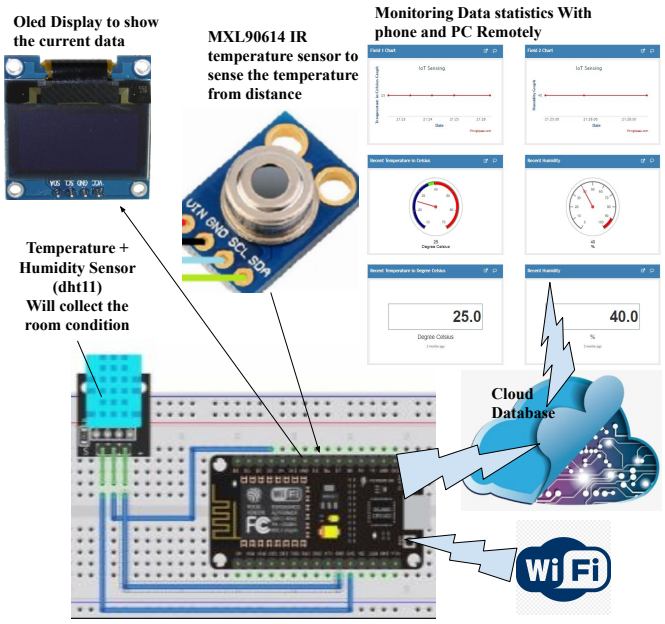


Fig. 5. Circuit Concept of Data Sensing and Sending Part

backward. According to the algorithm, when it detects three finger then the robot move to its right side and after detecting four fingers the robot moves to its left side. Finally when the camera detects five fingers of hand then the robot stops its movement. We use python programming language to detect fingers. Figure 6 shows detection of one finger simulation.

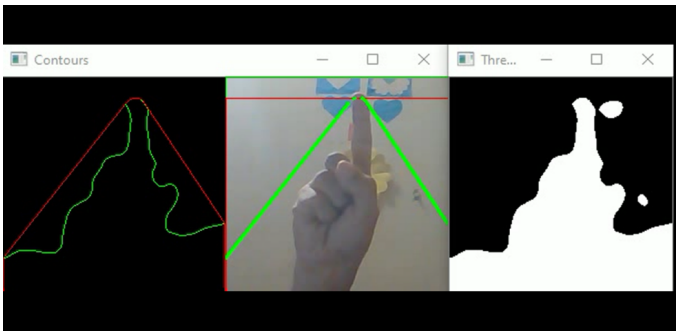


Fig. 6. Detect One Finger

The proposed system can detect two finger successfully. Figure 7 shows our system detect two fingers successfully.

Our system can detect two finger successfully. Figure 8 shows our system detect 3 fingers successfully.

The system can also detect two finger successfully. Figure 9 shows our system detect 4 fingers successfully.

Finally, our system detects five fingers form human hand. Figure 10 shows our system can detect 5 fingers accurately.

The system shows the collected data in the ThingSpeak server (Location: https://thingspeak.com/channels/739817). Figure 11 shows Collected data visualisation in cloud server.

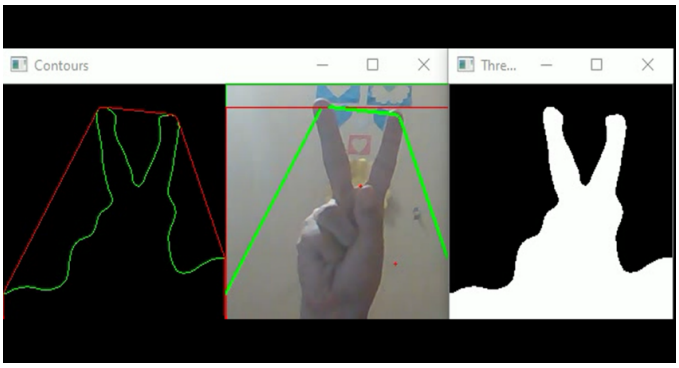


Fig. 7. Detect Two Fingers

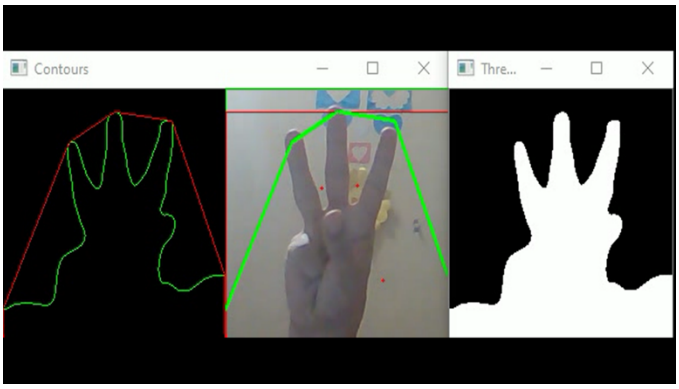


Fig. 8. Detect Three Finger

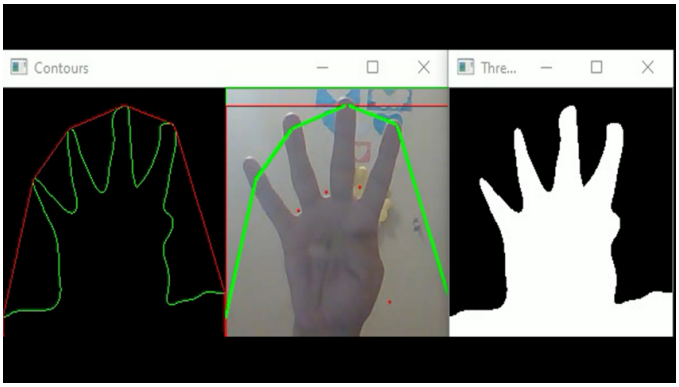


Fig. 9. Detect Four Finger

A. Prototype Output

Final view of the Prototype Robot is mentioned in Figure 12, Figure 13 and Figure 14.

B. Obtained Features and Results

We tested our system for posture detection, Bluetooth controlling and data sending to cloud server. We got desired results. The features obtained from the prototype are as follows:

1) The robot is able to recognize the posture and fingers of a patient and move towards the patient. It showed 95% success rate in 500 tests.

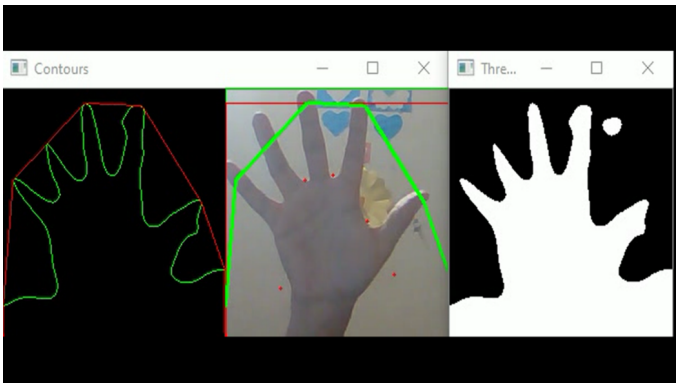


Fig. 10. Detect Five Finger

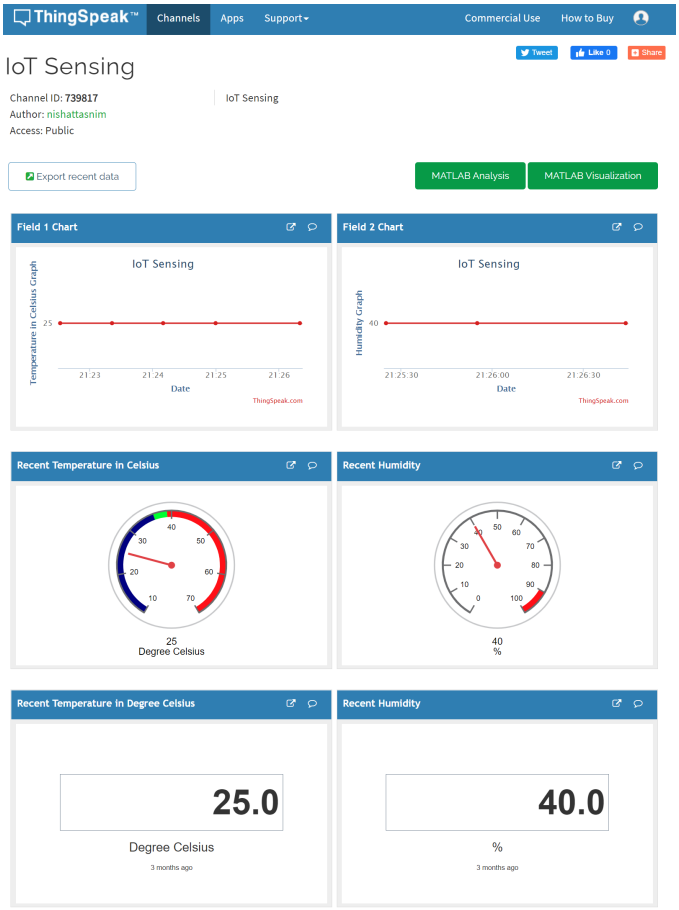


Fig. 11. Collected Data Visualisation in Cloud Server

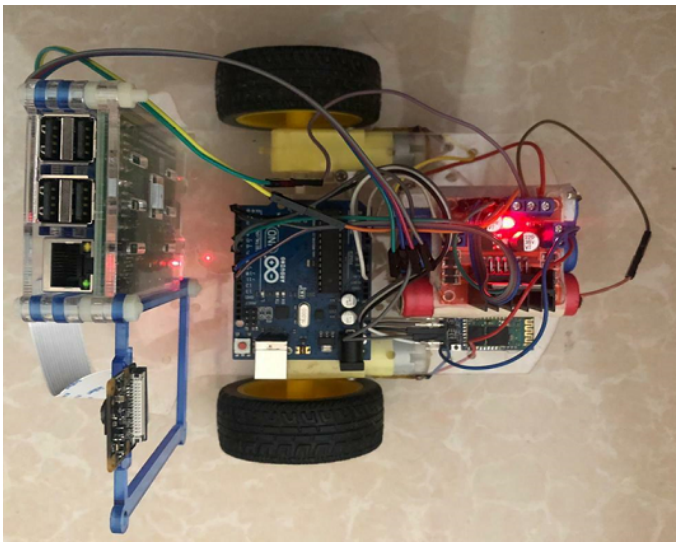


Fig. 12. Final Look (Side 1)

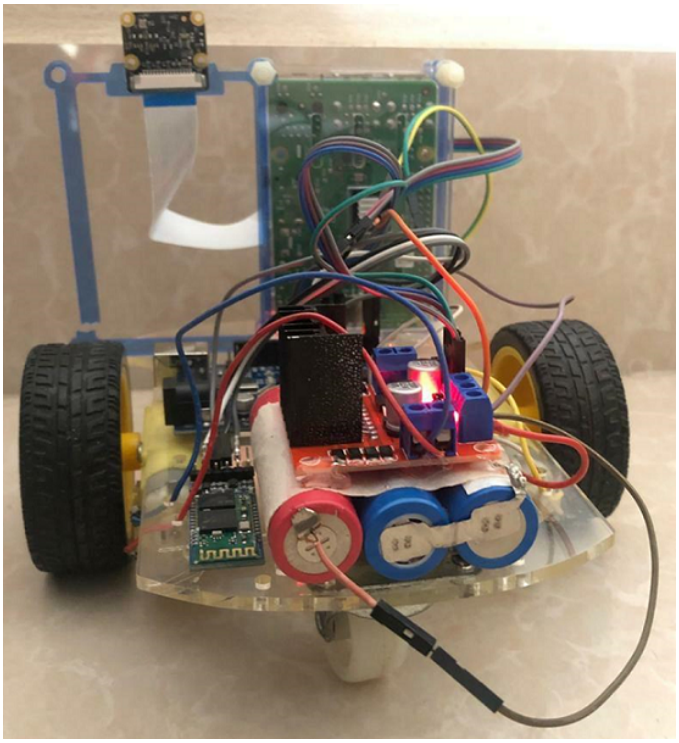


Fig. 13. Final Look (Side 2)

2) The robot can move 360 degree by following the instructions from posture. It showed 95% success rate in 500 tests.
3) The robot also works using Bluetooth sensor. It showed 98% success rate in 500 tests.
4) It can also make movement by the instructions from mobile phones via Bluetooth with 97% success rate in 500 tests.
5) The robot successfully collected temperature and humidity data from people and environment and showed in oled display with 96% success rate in 500 tests.
6) The robot successfully sent the collected data to cloud database with 94% success rate in 500 tests.
7) The system is cost effective. The prototype costs less than 100 USD. Where some existing systems may cost more than 500 USD without having all of the features obtained in this work.

C. Limitations

The system has some limitations also which can be mitigated in future development. The limitations we have found

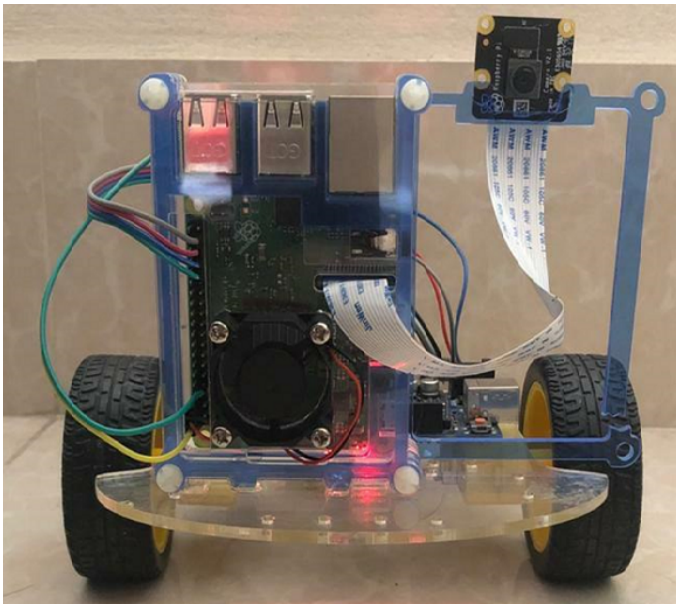


Fig. 14. Final Look (Side 3)

are as follows:

1) The system is not water proof. Water may damage the full or partial system.
2) For making the system low cost we used low cost sensors that may make incorrect results some time.
3) The robot should be maintained carefully. Otherwise it can be broken.

D. Discussion

The main objective of this paper is to operate a robot using hand gestures or a mobile phone through Bluetooth and after then, collect data and store them on a cloud server. The process of making hand gestures as input has been performed through an embedded pi camera, processed to recognize correctly as well as respond accordingly. The whole process takes on an average of 1500 milliseconds at the rate of 97% accuracy. Alam et al.[21] designed hand gesture-controlled robot using MPU 6050 gyroscope module and Arduino nano with an accuracy rate of 93.8%. Maharani et al.[24] performed 1080 number of tests in total by 6 different people with four gestures (forward, right, left, and stop), from three distances (2m, 3m, 4m), and at three slopes position (450, 00, -450). According to them, SVM showed superior results than K-means clustering with 95.15% recognition accuracy in 10ms. Hand gestures are used to control the home appliances[22] with the accuracy rate 87% in 3 meter distance. They utilized a Kinect sensor to take hand gestures with three-hand states(open, close, and lasso), and processing is performed through Raspberry Pi. Meanwhile, Xing et al.[38] achieved 83.23% accuracy using a little modified CNN to surface electromyographic signal for hand gestures recognition. Hence, it is evident that our proposed work provides superior the result compared to other existing methods summarised in table I.

Our robot also contains a non-contact temperature sensor MXL90614 which can detect Covid-19 affected patient's body temperature at a distance and send it to the cloud without wearing any devices in the body like the work of [29]. These above discussion shows the advantages of this work.

TABLE I. COMPARISON WITH STATE-OF-THE ART METHOD OF HAND POSTURE RECOGNITION

Method	Accuracy
Alam et al.[21]	93.8%
Fakhrurroja et al.[22]	87%
Nazzi et al.[38]	95.01%
Maharani et al.[24]	95.15%
Our Proposed Model	**97%**

V. CONCLUSION

In the modern era, robots are engaged in a lot of sectors. This work implemented an IoT-based posture recognizer remote sensing robot. Hospital patients can call the robot with hand posture and smartphone via Bluetooth; then the robot can go to the patient by following posture or Bluetooth command. Then it can collect data with sensors and send them to a cloud database. Therefore, any disabled or virus-affected person can control it and be monitored with this system remotely without affecting any healthy people. We got around 95-97% success rate among all the features. In future work machine learning features can be added to this robot to predict the patient and environmental conditions.

ACKNOWLEDGMENT

This research is partially supported through the Australian Research Council Discovery Project: DP190100314, "Re-Engineering Enterprise Systems for Microservices in the Cloud".

REFERENCES

[1] Atia Sultana, Md. Abul Hasan, and Tajim Md. Niamat Ullah Akhund, "An approach to Create IOT based Automated Smart Farming System for Paddy Cultivation," 2019. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.13443.04647

[2] T. M. N. U. Akhund, S. R. Snigdha, M. S. Reza, N. T. Newaz, M. Saifuzzaman, and M. R. Rashel, "Self-powered IoT-Based Design for Multi-purpose Smart Poultry Farm," in *Information and Communication Technology for Intelligent Systems*, T. Senjyu, P. N. Mahalle, T. Perumal, and A. Joshi, Eds. Singapore: Springer Singapore, 2021, vol. 196, pp. 43–51. [Online]. Available: http://link.springer.com/10.1007/978-981-15-7062-9_5

[3] T. M. N. U. Akhund, M. A. B. Siddik, M. R. Hossain, M. M. Rahman, N. T. Newaz, and M. Saifuzzaman, "IoT Waiter Bot: A Low Cost IoT based Multi Functioned Robot for Restaurants," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. Noida, India: IEEE, Jun. 2020, pp. 1174–1178. [Online]. Available: https://ieeexplore.ieee.org/document/9197920/

[4] T. M. Niamat Ullah Akhund, M. J. N. Mahi, A. N. M. Hasnat Tanvir, M. Mahmud, and M. S. Kaiser, "ADEPTNESS: Alzheimer's Disease Patient Management System Using Pervasive Sensors - Early Prototype and Preliminary Results," in *Brain Informatics*, S. Wang, V. Yamamoto, J. Su, Y. Yang, E. Jones, L. Iasemidis, and T. Mitchell, Eds. Cham: Springer International Publishing, 2018, vol. 11309, pp. 413–422. [Online]. Available: http://link.springer.com/10.1007/978-3-030-05587-5_39

[5] Tajim Md. Niamat Ullah Akhund, "Study and Implementation of Multi-Purpose IoT Nurse-BoT," 2019. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.14674.09924

[6] T. M. Niamat Ullah Akhund, W. B. Jyoty, M. A. B. Siddik, N. T. Newaz, S. A. Al Wahid, and M. M. Sarker, "IoT Based Low-Cost Robotic Agent Design for Disabled and Covid-19 Virus Affected People," in *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. London, United Kingdom: IEEE, Jul. 2020, pp. 23–26. [Online]. Available: https://ieeexplore.ieee.org/document/9210389/

[7] M. Oudah, A. Al-Naji, and J. Chahl, "Hand Gesture Recognition Based on Computer Vision: A Review of Techniques," *Journal of Imaging*, vol. 6, no. 8, p. 73, Jul. 2020. [Online]. Available: https://www.mdpi.com/2313-433X/6/8/73

[8] A. Boyali, N. Hashimoto, and O. Matsumoto, "Hand posture and gesture recognition using MYO armband and spectral collaborative representation based classification," in *2015 IEEE 4th Global Conference on Consumer Electronics (GCCE)*. Osaka, Japan: IEEE, Oct. 2015, pp. 200–201. [Online]. Available: http://ieeexplore.ieee.org/document/7398619/

[9] V.-T. Nguyen, T.-L. Le, T.-H. Tran, R. Mullot, and V. Courboulay, "Hand Posture Recognition Using Kernel Descriptor," *Procedia Computer Science*, vol. 39, pp. 154–157, 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050914014410

[10] T. M. Akhund, N. Ullah, N. T. Newaz, M. Rakib Hossain, and M. Shamim Kaiser, "Low-cost smartphone-controlled remote sensing iot robot," in *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*. Springer, 2021, pp. 569–576.

[11] T. Akhund, I. A. Sagar, and M. M. Sarker, "Remote temperature sensing line following robot with bluetooth data sending capability," in *International Conference on Recent Advances in Mathematical and Physical Sciences (ICRAMPS)*, 2018.

[12] T. M. Akhund, N. Ullah, G. Roy, A. Adhikary, A. Alam, N. T. Newaz, M. Rana Rashel, M. Abu Yousuf *et al.*, "Snappy wheelchair: An iot-based flex controlled robotic wheel chair for disabled people," in *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*. Springer, 2021, pp. 803–812.

[13] T. Akhund, N. T. Newaz, and M. M. Sarker, "Posture recognizer robot with remote sensing for virus invaded area people," *Journal of Information Technology (JIT)*, vol. 9, pp. 1–6, 2020.

[14] F. I. Suny, M. R. Fahim, M. Rahman, N. T. Newaz, T. M. Akhund, N. Ullah *et al.*, "Iot past, present, and future a literary survey," in *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*. Springer, 2021, pp. 393–402.

[15] N. T. Newaz, M. R. Haque, T. M. N. U. Akhund, T. Khatun, M. Biswas, and M. A. Yousuf, "Iot security perspectives and probable solution," in *2021 Fifth World Conference on Smart Trends in Systems Security and Sustainability (WorldS4)*. IEEE, 2021, pp. 81–86.

[16] A. H. Himel, F. A. Boby, S. Saba, T. M. Akhund, N. Ullah, and K. Ali, "Contribution of robotics in medical applications a literary survey," in *Intelligent Sustainable Systems*. Springer, 2022, pp. 247–255.

[17] M. Suny, F. Islam, T. Khatun, Z. Zaman, M. Fahim, M. Roshed, M. Islam, R. Jesmin, T. M. Akhund, N. Ullah *et al.*, "Smart agricultural system using iot," in *Intelligent Sustainable Systems*. Springer, 2022, pp. 73–82.

[18] M. R. Rashel, M. Islam, S. Sultana, M. Ahmed, T. M. Akhund, N. Ullah, J. N. Sikta *et al.*, "Internet of things platform for advantageous renewable energy generation," in *Proceedings of International Conference on Advanced Computing Applications*. Springer, 2022, pp. 107–117.

[19] M. M. Sarker and T. M. N. U. Akhund, "The roadmap to the electronic voting system development: a literature review," *International Journal of Advanced Engineering, Management and Science*, vol. 2, no. 5, p. 239465, 2016.

[20] M. M. Sarker, M. A. I. Shah, T. Akhund, and M. S. Uddin, "An approach of automated electronic voting management system for bangladesh using biometric fingerprint," *International Journal of Advanced Engineering Research and Science*, vol. 3, no. 11, p. 236907, 2016.

[21] M. Alam and M. A. Yousuf, "Designing and Implementation of a Wireless Gesture Controlled Robot for Disabled and Elderly People," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. Cox'sBazar, Bangladesh: IEEE, Feb. 2019, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8679290/

[22] H. Fakhrurroja, A. Abdillah, U. Nadiya, and M. Arifin, "Hand State Combination as Gesture Recognition using Kinect v2 Sensor for Smart Home Control Systems," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*. BALI, Indonesia: IEEE, Nov. 2019, pp. 74–78. [Online]. Available: https://ieeexplore.ieee.org/document/8980390/

[23] L. Chen, J. Fu, Y. Wu, H. Li, and B. Zheng, "Hand Gesture Recognition Using Compact CNN via Surface Electromyography Signals," *Sensors*, vol. 20, no. 3, p. 672, Jan. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/3/672

[24] D. A. Maharani, H. Fakhrurroja, Riyanto, and C. Machbub, "Hand gesture recognition using K-means clustering and Support Vector Machine," in *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. Penang: IEEE, Apr. 2018, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8405435/

[25] M. Meghana, C. Usha Kumari, J. Sthuthi Priya, P. Mrinal, K. Abhinav Venkat Sai, S. Prashanth Reddy, K. Vikranth, T. Santosh Kumar, and A. Kumar Panigrahy, "Hand gesture recognition and voice controlled robot," *Materials Today: Proceedings*, vol. 33, pp. 4121–4123, 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2214785320350343

[26] A. Ali A. and R. Sarah A., "Python-based Raspberry Pi for Hand Gesture Recognition," *International Journal of Computer Applications*, vol. 173, no. 4, pp. 18–24, Sep. 2017. [Online]. Available: http://www.ijcaonline.org/archives/volume173/number4/abed-2017-ijca-915285.pdf

[27] A. V. and R. R., "A Deep Convolutional Neural Network Approach for Static Hand Gesture Recognition," *Procedia Computer Science*, vol. 171, pp. 2353–2361, 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050920312473

[28] C. Chansri, J. Srinonchat, E. G. Lim, and K. L. Man, "Low Cost Hand Gesture Control in Complex Environment Using Raspberry Pi," in *2019 International SoC Design Conference (ISOCC)*. Jeju, Korea (South): IEEE, Oct. 2019, pp. 186–187. [Online]. Available: https://ieeexplore.ieee.org/document/9027669/

[29] M. S. Mondal, K. Roy, and S. Sarkar, "Design and Development of Wearable Remote Temperature Monitoring Device for Smart Tracking of COVID-19 Fever," *SSRN Electronic Journal*, 2020. [Online]. Available: https://www.ssrn.com/abstract=3735919

[30] T. M. Akhund, N. Ullah, N. T. Newaz, Z. Zaman, A. Sultana, A. Barros, and M. Whaiduzzaman, "Iot-based low-cost automated irrigation system for smart farming," in *Intelligent Sustainable Systems*. Springer, 2022, pp. 83–91.

[31] M. Whaiduzzaman, A. Barros, A. R. Shovon, M. R. Hossain, and C. Fidge, "A resilient fog-iot framework for seamless microservice execution," in *2021 IEEE International Conference on Services Computing (SCC)*. IEEE, 2021, pp. 213–221.

[32] R. Hossen, M. Whaiduzzaman, M. N. Uddin, M. J. Islam, N. Faruqui, A. Barros, M. Sookhak, and M. J. N. Mahi, "Bdps: An efficient spark-based big data processing scheme for cloud fog-iot orchestration," *Information*, vol. 12, no. 12, p. 517, 2021.

[33] M. Whaiduzzaman, M. J. N. Mahi, A. Barros, M. I. Khalil, C. Fidge, and R. Buyya, "Bfim: Performance measurement of a blockchain based hierarchical tree layered fog-iot microservice architecture," *IEEE Access*, vol. 9, pp. 106 655–106 674, 2021.

[34] N. Faruqui, M. A. Yousuf, M. Whaiduzzaman, A. Azad, A. Barros, and M. A. Moni, "Lungnet: A hybrid deep-cnn model for lung cancer diagnosis using ct and wearable sensor-based medical iot data," *Computers in Biology and Medicine*, vol. 139, p. 104961, 2021.

[35] M. R. Hossain, M. Whaiduzzaman, A. Barros, S. R. Tuly, M. J. N. Mahi, S. Roy, C. Fidge, and R. Buyya, "A scheduling-based dynamic fog computing framework for augmenting resource utilization," *Simulation Modelling Practice and Theory*, vol. 111, p. 102336, 2021.

[36] M. Mahi, J. Nayeen, K. M. Hossain, M. Biswas, and M. Whaiduzzaman, "Sentrac: A novel real time sentiment analysis approach through twitter cloud environment," in *Advances in Electrical and Computer Technologies*. Springer, 2020, pp. 21–32.

[37] A. M. R. Mazumder, K. A. Uddin, N. Arbe, L. Jahan, and M. Whaiduzzaman, "Dynamic task scheduling algorithms in cloud computing," in *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2019, pp. 1280–1286.

[38] K. Xing, Z. Ding, S. Jiang, X. Ma, K. Yang, C. Yang, X. Li, and F. Jiang, "Hand Gesture Recognition Based on Deep Learning Method," in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. Guangzhou: IEEE, Jun. 2018, pp. 542–546. [Online]. Available: https://ieeexplore.ieee.org/document/8411908/

Deepfakes on Retinal Images using GAN

Yalamanchili Salini
School of Computer Science & Engineering
VIT-AP University, Amaravati
Andhra Pradesh, India

Dr J HariKiran
School of Computer Science & Engineering
VIT-AP University,Amaravati
Andhra Pradesh, India

***Abstract*—In Deep Learning (DL), Generative Adversarial Networks (GAN) are a popular technique for generating synthetic images, which require extensive and balanced datasets to train. These Artificial Intelligence systems can produce synthetic images that seem authentic, known as Deep Fakes. At present, data-driven approaches to classifying medical images are prevalent. However, most medical data is inaccessible to general researchers due to standard consent forms that restrict research to medical journals or education. Our study focuses on GANs, which can create artificial fundus images that can be indistinguishable from actual fundus images. Before using these fake images, it is essential to investigate privacy concerns and hallucinations thoroughly. As well as, reviewing the current applications and limitations of GANs is very important. In this work, we present the Cycle-GAN framework, a new GAN network for medical imaging that focuses on the generation and segmentation of retinal fundus images.DRIVE retinal fundus image dataset is used to evaluate the proposed model's performance and achieved an accuracy of 98.19%.**

***Keywords*—DeepFakes; deep learning; retinal fundus image synthesis; segmentation; generative adversarial network (GAN); variational autoencoder (VAE)**

I. Introduction

An eye's retina is a sensitive membrane responsible for vision.As shown in Figure 1, three primary anatomical components are the Optic Disc, Macula, and Blood Vessels.

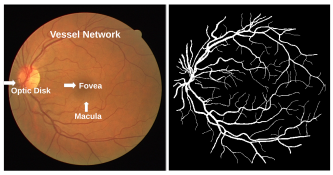


Fig. 1. Picture of the Retina on the Left; The Segmented Image on the Right.

To categorize the GAN's working capability, we divided them into seven categories: synthesis, segmentation, reconstruction, detection, de-noising, registration, and classification. The use of GANs has been studied across many different imaging modalities, including MRI (magnetic resonance imaging), CT (computed tomography), OCT (optical coherence tomography), chest X-rays, dermoscopy, ultrasound, PET, and microscopy.

A classic area of study in computer vision is image classification. A large, well-balanced dataset is frequently needed for training deep neural networks. However, because of the unbalanced dataset, most networks' performance will suffer while classifying medical images. Moreover, collecting pathological instances takes time in the domain of medical images. The ideal option is to create new, high-quality, diverse photographs of minority classes[1].

Artificial intelligence (AI) has gained popularity in recent years for use in medical imaging jobs [2]. However, even while medical data sets are more widely available, most of them only apply to certain medical diseases, and collecting data for machine learning methods is still tricky [3,4]. Some initiatives have focused on adding to the existing data to get beyond this obstacle. Numerous techniques for data augmentation have been proposed in this regard. Despite this, only minor adjustments, such as overfitting in learning processes or geometric modifications, have been made to meet the urgent requirement to provide data sets more meaningful [5,6]. However, considerable improvement has been accomplished by introducing synthetic data augmentation to extend training sets. For example, synthetic data can present novel photos to existing data sets. It might contribute to increased diversity within a dataset and, eventually, to more robust machine learning algorithms if such a strategy is adopted.

To achieve the mentioned improvements: 1) GANs exploit density ratio estimation in an indirect manner of supervision to maximize probability density over the data-generating distribution; 2) By discovering the latent distribution of high-dimensional data, GANs have improved the performance of visual feature extraction.

For all these Deepfakes comes into picture because Deepfakes have gained public attention for their sinister uses, but they have also investigated in several medical fields [7,8]. As ophthalmology has been at the forefront of the DL revolution, synthetic images can be used for various purposes, including fundus[9,10,11] and OCT. Several potential uses of GANs in ophthalmology have yet to be investigated, including how they can be applied to DL development and medical education [12,13] and the implications of their use for privacy regulations. This study had two goals:

- A GAN applied to synthetic images generated by using DRIVE database was tested to determine whether the machine could identify the authentic fundus images.
- In addition, GANs are being examined for their uses in ophthalmology, as well as their limitations.

The remaining portions of this paper take place in multiple sections—first, the related work regarding Image translation and Image synthesis is discussed in Section II Then, Sections III goes on with Materials & Methods for retinal image

generation. Next, the proposed network and its importance will discuss in Section IV Next, the experimentation findings take part in Section V where the segmentation's performance and execution time are concerned with existing techniques. Later on, concluding with a discussion in Section VI. Finally, Section VII contains a conclusion.

II. Related Work

Deep learning-based computer systems that assist in medical diagnostics are greatly interested. But because of restrictions on data access due to proprietary and privacy issues, these systems' development and improvement cannot be sped up by contributions from the general public [14]. For example, without the patient's consent, it might be challenging for medical personnel to publish most medical pictures [15]. Furthermore, the publicly accessible datasets frequently have an insufficient size and expert annotations, making them unsuitable for training data-hungry neural networks. As a result, only academics with access to private data can create these systems, which restricts the development and potential of this area of study.

A. GAN & VAE

In addition to GAN, Variational Autoencoder(VAE) is another family of deep generative models that should investigate for medical imaging tasks. Latent (random) vectors are the input for GAN. However, one must carefully modify the GAN output to create synthetic images with the required characteristics. To deal with this issue, VAE had introduced. An encoder and a decoder are the two components of a VAE. Utilizing multilayer convolutional neural networks, the encoder turns input images into latent vectors of random variables with corresponding mean and standard deviations.

VAE, unlike GAN, starts with samples selected from the latent vector associated with the input and then sends them to the decoder for reconstruction. Thus, we can manipulate VAE directly to create specific synthetic output images for clear input photos. However, due to the loss function of the mean square error, the output of the VAE could appear hazy. Combining the advantages of VAE and GAN creates an adversarial network for similarity measures to address this problem. The application of VAE in medical imaging is quite innovative [16,17] and needs further investigation to process retinal images.

B. Image-to-Image Translation

In picture-to-image translation, an altered version of an existing image is created synthetically. Therefore, a sizable dataset of matched instances is often needed when training a model for image-to-image translation. For which a paired sample dataset is traditionally required to prepare an image-to-image translation model. In other words, a sizable dataset with several examples of modified versions of the input image X that can be utilised as the intended output image Y. These datasets, particularly in the medical field, are time-consuming, expensive, and sometimes impossible to compile. The image-to-image translation framework can be applied to a variety of computer vision problems, including image super-resolution [18], image inpainting [19], and style transfer [20]. It is possible to employ both supervised and unsupervised methods [21,22,23].

C. Retinal Image Synthesis

Surgical simulations using an anatomic model of the eye and surrounding face were one of the first applications of retinal image synthesis. Nevertheless, the segmentation module's performance heavily influences the quality of the generated images. To reduce the requirement for annotated samples and to improve the representativeness (for example, the variability) of synthesized images [24], a generative adversarial approach is used in conjunction with a style transfer algorithm. Recent implementations like the retinal background and fovea have been modelled using a dictionary of small images without vessels [25]. In addition, it's an idea that training a segmentation network with authentic retinal images combined with synthesized ones leads to better segmentation results.

D. GAN's on Retinal Image Synthesis: Present Status

GANs have shown the ability to produce impressively realistic synthetic medical images. This section describes existing work on GANs for synthesising coloured retinal fundus images [26,27,28,29,30,31]. (Table I)

III. Materials & Methods

A. Dataset

The DRIVE dataset initially consisted of 40 photos, but we expanded it to 120 images, using 125 for training, 55 for validation, and 20 for testing. This image used with a field view of 45 degrees and a dimension of 565 x 584 pixels. It has 540 pixels in diameter and a FOV of 540 pixels. As seen in Fig 2, each image in the DRIVE dataset has a mask to aid in identifying the field of view (FOV) region.

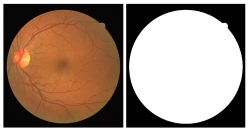


Fig. 2. DRIVE Database Sample (a) Original Image. (b) Mask Image.

B. Image Preparation

A black-and-white retinal vasculature map was created for each image using a U-Net trained on 154 photos from the DRIVE database [32].The unaided eye cannot detect pigmentation and choroidal blood vessel patterns on vessel maps, so information about them is removed. In addition, a circular mask with black background was placed on all retinal images with suitable vascular maps to create photos of the synthetic retinal fundus images.

TABLE I. LIST OF ARTICLES ON THE CREATION OF COLOURED RETINAL IMAGES

References	DataSets	Methods	Validation
26	i-ROP	PGANs	Segementation and Latent space espression
27	Messidor	cGAN(Pix2Pix)	ISC,Qv
28	Messidor	AAE and cGAN(Pix2Pix)	Segementation and ISC
29	Messidor	cGAN(Pix2Pix)	Segementation and SSIM
30	Drive, Stare and Style references	cGAN(Pix2Pix) and Style transfer	Segementation
31	Drive , Stare, HRF and Style references	cGAN(Pix2Pix) and Style transfer	Segementation and SSIM

C. Why GANs?

GANs are deployed and used for artificial data augmentation. GANs work through the creation of synthetic pictures while simultaneously learning to distinguish between them as actual pictures see Fig 3. In addition to their use in ophthalmology, GANs are helpful in molecular oncology imaging and generated positron emission tomography (PET) pictures [33]. Even though present radiology applications attempt to aid in the diagnosis, human perception has not yet been used in this situation to assess the quality of GAN created synthetic data. In several instances, using GAN improves medical imaging by creating fresh retinal pictures from data consisting of pairs of retinal vascular trees [34]. Generator loss function and Discriminator classification information about generated images are depicted as well as Convolutional neural networks (CNNs) are standard tools for categorizing images and returning a scalar to represent the realness of the input pictures.

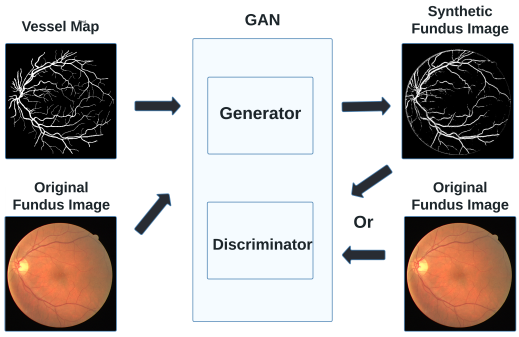


Fig. 3. Generative Adversarial Network Training, pix2pix.

D. U-Net

In order to generate a wider range of realistic images, we developed a pipeline instead of CNN based on this we trained a U-Net segmentation network with our synthetic data to generate a segmentation mask from a photorealistic medical image to assess the credibility of the data. The u-net design, explicitly created for biomedical images, is descended from the auto encoder architecture, which uses unsupervised learning for dimensionality reduction. The u-net is particularly helpful for biomedical applications because it lacks completely connected layers, has no restrictions on the size of input images and permits a substantially higher number of feature channels than a conventional CNN [35]. The decoding procedure also concatenates the receptive fields before and after convolution. By doing this, the network can use both the up-convolutional and initial properties. To determine the accuracy of the GAN, 4282 image pairs were trained for 200 epochs. Following this, synthetic retinal fundus images were created using all the retinal vascular maps from the test data. It is one of the key advantages of GANs that they can produce much larger datasets than the initial ones see Fig 4.

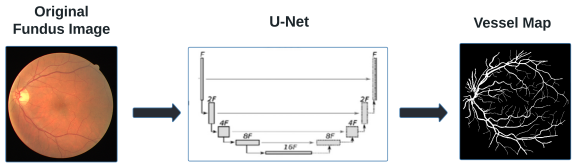


Fig. 4. U-Net Segmentation Network

E. Segmentation

Machine learning involves segmenting images into appropriate sections. Fundus pictures with low contrast, complicated, and compound characteristics must be meticulously segmented to separate retinal vessels from one another. Deep learning systems are capable of identifying vessels against backgrounds accurately. This method, however, did not factor in ambiguous vessels, resulting in inaccurate estimates of vascular calibre biomarkers, such as tortuosity, length-to-diameter ratios, branching angles, and fractal dimensions. The proposed architecture uses long and short skip connections along with U-Net to address the abovementioned problem. Segmenting retinal vessels and looking for anomalies in the retinal subspace requires an exact technique. In recent years, several supervised and unsupervised algorithms have been proposed to segment retinal vessels. However, manual feature extraction is necessary for training with supervised approaches for different applications [36], [37].In below we can see the workflow of supervised and unsupervised algorithms.

- A minimization function is used over the tuning process to determine which separation between the vascular and background classes is the most effective. Fig 5 displays a typical unsupervised learning algorithm workflow.

- In supervised approaches, the segmentation algorithm must learn the vessel segmentation rule by studying the images manually labelled by professionals. Fig 6 depicts the workflow of a typical supervised technique.

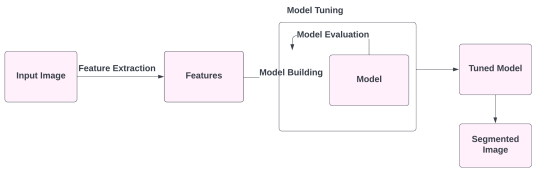


Fig. 5. Unsupervised Learning Algorithm Workflow.

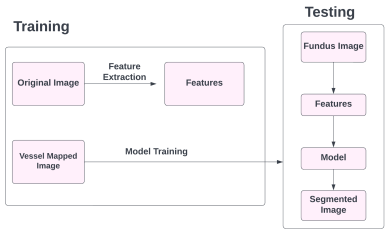


Fig. 6. Supervised Learning Algorithm Workflow.

For getting segmented image initially in the first set, 577,649 pixels (12.7 percent) are marked as vessels, while 556,532 pixels (12.3 percent) are marked as vessels in the testing set, which is segmented twice from the training set [38]. See Fig 7 and Fig 8 for detailed view of DRIVE dataset segmentation and masking. Our results are comparable to those achieved by state-of-the-art methods using U-Net implementation to Cycle-GAN Network.

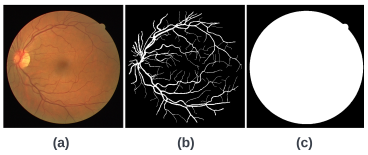


Fig. 7. A Sample from the DRIVE Dataset. (a) Training Image, (b) Vessel Segmented of the Training Image, and (c) Mask of the Image.

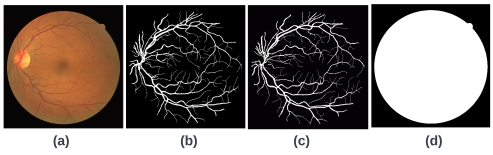


Fig. 8. (a) A Sample Test Image from the DRIVE Dataset. (b) The First Segmentation (c) The Second Segmentation (d) The Mask Image.

F. PreProcessing

During this stage, the retinal image quality is enhanced by separating vessels from the backdrop for achieving segmentation of vessels accuracy.The recommended network ensures that the retinal vascular tree can be segmented more effectively. Therefore, the trained model of the suggested network serves as the foundation for our method for retinal vascular segmentation, and its processing pipeline, as shown in Fig 9. It should note that using the DL network to segment a complete image may produce unreliable results. For the suggested neural network to focus, it is necessary to crop photos into patches.We will repeat this process in testing to produce segmented patches using the trained model. The segmented vessel tree is then produced by merging the segmented patches during the post-processing stage as shown in Fig 9.

IV. Proposed Work

A. Cycle-GAN: General Pipeline

Any model should be able to identify the underlying relationship between the two domains and extract distinctive features from each field for image transformation between them. Cycle-GAN is nominated to offer these guidelines[39]. The finding in (1) briefs a mapping between domain X and domain Y, and vice versa, the system essentially merges two GANs. A generator G: X Y trained by discriminator DY and a generator F: Y X trained by discriminator DX create a structure shown in Fig 10.

$$minmax(D,G) = E_x{-}Pdata(x)[logD(x)]{+}E_x{-}Py[1{-}D(G(y)))] \quad (1)$$

B. Loss Function

No paired data is available for CycleGAN training, so the input X and the target Y pair are not guaranteed to be meaningful. Thus, we propose the Cycle Consistency loss to ensure the network learns the correct mapping. Both discriminator loss and generator loss are similar to those used in pix2pix.

A cycle consistency refers to a close match between the input and the output. For Example, when we talk about NLP translations, the resulting sentence should be the same as the original sentence when translating from English to Telugu and then back to English. As a result of cycle consistency loss as specified in (2) and (3) :

- X image information is passed to generator G, which produces image Y1.
- A cycled image Y1 is generated by passing generated image F through generator X1.
- Between X and X1, we calculate the mean absolute error. In the Figure 11, generator G is responsible for converting image X into image Y. If you feed image Y to generator G, and the output would be the image Y itself or something close.

$$Forwardcycleloss : X \rightarrow G(X) \rightarrow F(G(X))\tilde{}X1 \quad (2)$$

$$Backwardcycleloss : Y \rightarrow F(Y) \rightarrow G(F(Y))\tilde{}Y1 \quad (3)$$

C. Image Generation

The validation dataset examined images created from retinal vessel maps manually after training the GAN for 100 epochs on 120 pairs of images.Using all vessel maps, produced a synthetic retinal fundus image from the test dataset, see Fig 12 how the synthetic image looks by using proposed network.

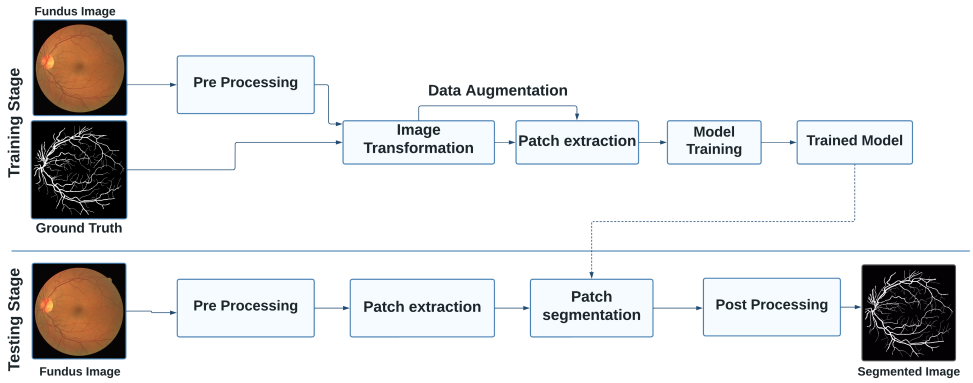


Fig. 9. Preprocess Functioning.

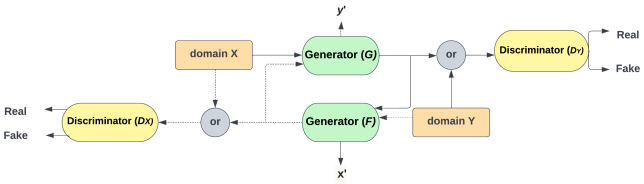


Fig. 10. Image Transformation using Cycle-GAN.

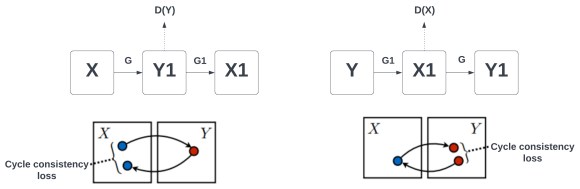


Fig. 11. Cycle Consistency Loss.

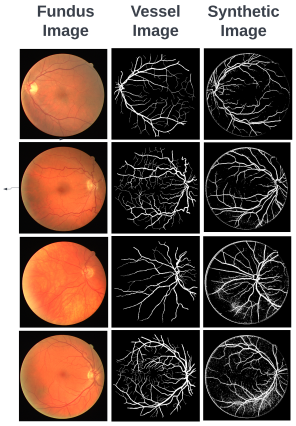


Fig. 12. A Trained U-Net is used to Segment Authentic Retinal Fundus Images (Left) into Associated Vessel Maps (Middle). In Order to Create the Artificial Retinal Fundus Images, We used pix2pixHD, a Newly Developed Implementation of a Generative Adversarial Network (GAN).

D. PostProcessing

A segmented blood vessel image is created by merging all segmented patches.As a result, the offered patches are gathered and reduced in size for cropping. These patches are then replicated in the appropriate order, depending on the image size for cropping[40]. To remove the white pixels surrounding the retina, the mask of the used picture is placed on the combined image. Then, noise is removed using the morphological transformation ”erosion” utilising an ellipse structural element of size 2*2.

V. Experiments and Analysis

In this section, it is explained about the Parameter Settings in Section A. Later on, the evaluation principle is described in Section B, where the method is configured and put into practice. Then, using a retinal image dataset, Image classification is provided in Section C. Finally, we will see execution time measures in Section D.

A. Parameter Setting

Segmentation performance is achieved by training the suggested network with parameters selected experimentally or by consulting recent works. Experimentally, we determine the learning rate, the optimizer algorithm, the weight initialization method, and the epoch number [41]. First, we train one model without changing the parameters. Next, we pick the value with the highest segmentation rate.

B. Evaluation Principle and Metrics

We advise comparing the segmentation findings with manual segmentation by a skilled medical professional. Each pixel is defined as True Positives (T_P), True Negatives (T_N), False Positives (F_P), or False Negatives for the evaluation (F_N). Pixels correctly identified as background or vessels are expressed as T_P and T_N, respectively. As opposed, F_P and F_N represent pixels incorrectly identified as background or boats. A segmentation performance measure consists of Accuracy, Sensitivity, Specificity, and F1-Score. These metrics are the ones that are used most often to evaluate segmentation results. To classify

pixels as vessels Accuracy performance is calculated, while Sensitivity and Specificity represent the ability to categorize pixels as vessels and backgrounds. The Precision parameter specifies the percentage of correctly classified background and vessel pixels among all correctly classified background and vessel pixels. As shown in Table II the suggested method employs the following performance metrics.

Table III provides the performance metrics on DRIVE dataset where our method achives 98.19% accuracy in detecting segmented images.The obtained ROC curves and plots representation for the performance metrics is shown in below Figure 13, Figure14.

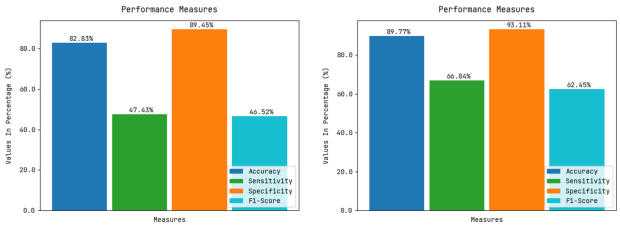


Fig. 13. Two Plots Measures for Test and Tain on DRIVE Dataset.

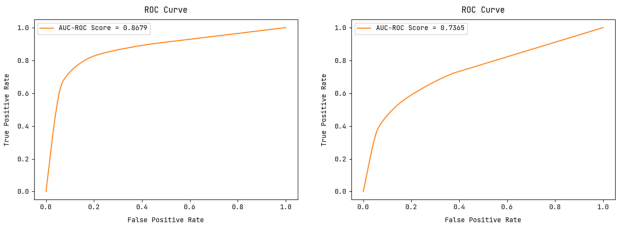


Fig. 14. Two Obtained ROC Curves on DRIVE Dataset.

TABLE II. PERFORMANCE METRICS

Metric	Elucidation
Accuracy	T_N+T_P/T_P+F_P+T_N+F_N
Sensitivity	T_P/(T_P+F_N)
Specificity	T_N/(T_N+F_P)
F1-Score	T_P/(T_P + F_P)

C. Image Classification

Using the high dimensional space, we can calculate the conditional probability, P(ai—aj), representing the similarity between two samples is shown in (4).

$$P(ai|aj) = \frac{exp\left(-\frac{-|ai-aj|^2}{2\sigma^2}\right)}{\Sigma_k \#1 exp\left(-\frac{-|ak-al|^2}{2\sigma^2}\right)} \quad (4)$$

50 actual and 50 synthetic photos with the same stage and illness distribution as the original dataset were uploaded, and runned ML programs to judge whether the photographs were natural or artificial. According to Figure 15 findings, most machine programs significantly distinguish between actual and artificial photographs.

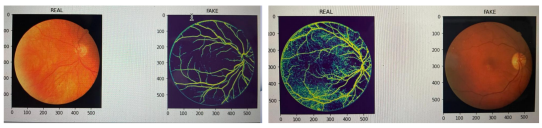


Fig. 15. Classfying Real and Fake Fundus Images.

D. Execution Time Measures

The proposed method is examined in this section for its processing performance. As shown in Table(IV),we propose calculating each image's execution period from the DRIVE dataset, respectively.Our analysis shows that despite the size of the image used, the computation values are too low for preprocessing, segmentation, and postprocessing.Then, we proposed evaluating the accuracy of the execution time compared to existing methods. Timing data is used in the evaluation for tarining the data. Because DRIVE is the most frequently used database, where values are provided in Table(V) , both metrics correspond to that database.

VI. DISCUSSION

In medical imaging, variations in illumination, noise, patterns, etc., result in a nonconvincing image produced by a GAN. A poorly defined vessel tree structure and dark spots show that the GAN can't distinguish complex systems. As a result, it can only identify colour, shape, and lighting features.There are many intricacies in medical images that must be accurately portrayed for the data to be useful for medical imaging. This lack of detail is unacceptable for medical image generation, as medical images contain many intricacies. By breaking down the complex task of generating medical ideas into hierarchical processes, our Cycle-GAN architecture improves the quality of synthetic images by using the below rules:

- In the first step of generating Images, GAN focuses on developing segmentation metrics by ignoring the realism of photos.
- Using this technique, in the second step, GAN concentrates only on generating the colour of an image, brightness of image, and texture of image based on the dimensions provided.

In addition, our proposed network generates more diverse photos than original dataset.With Fig 12, GAN is able to produce synthetic images by keeping general statistical classification of the real dataset.

The method of retinal image synthesis currently used for rebuilding the optic disc and fovea is quite adequate, but duplicate lesions with high fidelity is a challenge that requires further research. In addition, for quality validation, experts and ophthalmologists must assess the level of realism of generated images. As a result of this study, we were able to demonstrate the below points:

- That vessel maps of original retinal images obtained by ROP screening can yield realistic-appearing synthetic fundus images and

TABLE III. PERFORMANCE MEASURE VALIDATIONS ON DRIVE DATASET.

Database	No.of Epochs	Accuracy	Sensitivity	Specificity	F1-Score
DRIVE	25	82.83%	47.43%	89.45%	46.52%
	50	89.77%	66.84%	93.11%	62.45%
	75	98.19%	85.88%	99%	60.1%
	100	97.5%	73.18%	99.46%	89.17%

TABLE IV. DURATION PERIOD ON DRIVE DATASET IMAGES.

Metric	DRIVE Dataset
Preprocessing	0.026(s)
Segmentation	0.67(Time taken per patch)
Postprocessing	0.00347(s)
Time duration for each Image	0.7341(s)

TABLE V. COMPARISION TABLE OF TIME DURATION AND ACCURACY ON DRIVE DATASET.

References	Publication Year	Time Duration in(s)	Accuracy
34	2008	0.193	0.198
35	2012	6.8	0.9516
36	2018	0.421	0.943
37	2019	0.037	0.938
Our Method	2022	0. 69	0.9819

- That most of machine programs can distinguish natural from synthetic retinal images.Annotated data can be used to create innovative methods for analyzing retinal images or to enrich information in existing databases to create synthetic images that look as authentic as possible. Additionally, due to GAN's adaptability, they can be used to synthesize medical images using approaches used for retinal synthesis.

VII. CONCLUSION

The synthesis of retinal pictures using GANs has recently attracted more interest, and GANs have significantly developed in recent years. These tools can overcome restrictions like the scarcity of sizable annotated datasets and overcome the expensive expense of collecting high-quality medical data. However, the findings of GAN applications in the realm of medical imaging are still far from being practically applicable. The unique anatomy of a colour retinal fundus image must also be taken into consideration when generating synthetic retinal images in order to learn about a patient's health.

In this study, we present the Cycle-GAN framework, a new generative adversarial network for medical imaging that focuses on the generation and segmentation of retinal artery images. As a result, these artificial visuals appear realistic. DRIVE retinal fundus image dataset is used to evaluate the proposed model's performance and achieved an accuracy of 98.19%. We must focus on investigating datasets of various biomedical images for interaction, domain adaptation tasks, and segmentation of medical images in the future.

ACKNOWLEDGMENT

We would like to thank VIT-AP university for facilitating the resources required for conducting this research.

REFERENCES

[1] Huang, Gaofeng, and Amir Hossein Jafari. "Enhanced balancing GAN: Minority-class image generation." Neural Computing and Applications (2021): 1-10.

[2] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.

[3] Greenspan H, van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEETrans Med Imaging. 2016;35:1153–1159.

[4] Kazuhiro, Koshino, et al. "Generative adversarial networks for the creation of realistic artificial brain magnetic resonance images." Tomography 4.4 (2018): 159-163.

[5] Yang, Jiachen, et al. "MTD-Net: Learning to Detect Deepfakes Images by Multi-Scale Texture Difference." IEEE Transactions on Information Forensics and Security 16 (2021): 4234-4245.

[6] Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. ArXiv e-prints [Internet]. 2017 December 01, 2017.

[7] Greenspan H, van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEETrans Med Imaging. 2016;35:1153–1159.

[8] Crystal DT, Cuccolo NG, Ibrahim AMS, et al. Photographic and video deepfakes have arrived: how machine learning may influence plastic surgery. Plast Reconstr Surg. 2020;145: 1079e1086.

[9] Fallis D. The epistemic threat of deepfakes. Philos Technol. 2020:1e21 [Online ahead of print].

[10] Burlina P, Joshi N, Paul W, et al. Addressing artificial intel- ligence bias in retinal diagnostics. Transl Vis Sci Technol. 2021;10, 13e13.

[11] Burlina PM, Joshi N, Pacheco KD, et al. Assessment of deep generative models for highresolution synthetic retinal image generation of age-related macular degeneration. JAMA Oph- thalmol2019;137:258e264.

[12] Zhou Y, Wang B, He X, et al. DR-GAN: conditional gener- ative adversarial network for finegrained lesion synthesis on diabetic retinopathy images. IEEE J Biomed Health Inform. 2020, 1e1 [Online ahead of print].

[13] Costa P, Galdran A, Meyer MI, et al. End-to-end adversarial retinal image synthesis. IEEE Trans Med Imaging. 2018;37: 781e791.

[14] Secretary, HHS Office of the, and Office for Civil Rights (OCR). "Your Rights Under HIPAA." HHS.gov, US Department of Health and Human Services, 1 Feb. 2017.

[15] Cunniff, Christopher, et al. "Informed consent for medical photographs." Genetics in Medicine 2.6 (2000): 353-355.

[16] Chen, X., Pawlowski, N., Rajchl, M., Glocker, B., Konukoglu, E.: Deep generative models in the real-world: an open challenge from medical imaging. arXiv preprint arXiv:1806.05452 (2018)

[17] Tomczak, J.M., Welling, M.: Improving variational auto-encoders using house- holder flow. arXiv preprint arXiv:1611.09630 (2016)

[18] Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.

[19] Pathak, D.; Krähenbühl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.

[20] Gatys, L.A.; Ecker, A.S.; Bethge, M. A neural algorithm of artistic style. arXiv 2015, arXiv:1508.06576.

[21] Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv 2018,arXiv:1710.10196.

[22] Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.

[23] Chen, Q.; Koltun, V. Photographic Image Synthesis with Cascaded Refinement Networks. In Proceedings of the 2017 IEEE

[24] Zhao, He, et al. "Synthesizing retinal and neuronal images with generative adversarial nets." Medical image analysis 49 (2018): 14-26.

[25] Fiorini, S.; Biasi, M.D.; Ballerini, L.; Trucco, E.; Ruggeri, A. Automatic Generation of Synthetic Retinal Fundus Images. In SmartTools and Apps for Graphics—Eurographics Italian Chapter Conference; Giachetti, A., Ed.; The Eurographics Association: Cagliari,Italy, 2014.

[26] Beers, A., et al.: High-resolution medical image synthesis using progressively grown generative adversarial networks. arXiv preprint arXiv:1805.03144 (2018)

[27] Costa, P., et al.: Towards adversarial retinal image synthesis. arXiv preprint arXiv:1701.08974 (2017)

[28] Costa, P., et al.: End-to-end adversarial retinal image synthesis. IEEE Trans. Med. Imaging 37(3), 781–791 (2018)

[29] Guibas, J.T., Virdi, T.S., Li, P.S.: Synthetic medical images from dual generative adversarial networks. arXiv preprint arXiv:1709.01872 (2017)

[30] Iqbal, T., Ali, H.: Generative adversarial network for medical images (MI-GAN). J. Med. Syst. 42(11), 231 (2018)

[31] Zhao, H., Li, H., Maurer-Stroh, S., Cheng, L.: Synthesizing retinal and neuronal images with generative adversarial nets. Med. Image Anal. 49, 14–26 (2018)

[32] Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep con- volutional neural networks. JAMA Ophthalmol. 2018;136: 803e810.

[33] Ben-Cohen A, Klang E, Raskin SP, Amitai MM, Greenspan H. Virtual PET images from CT data using deep convolutional networks: initial results. In: Tsaftaris S, Gooya A, Frangi A, Prince J, eds. Simulation and Synthesis in Medical Imaging. New York: Springer, Cham. 2017. pp. 49–57.

[34] Costa P, Galdran A, Meyer MI, Niemeijer M, Abramoff M, Mendonca AM, Campilho A. End-to-end adversarial retinal image synthesis. IEEE Trans Med Imaging. 2018;37:781–791.

[35] Guibas, John T., Tejpal S. Virdi, and Peter S. Li. "Synthetic medical images from dual generative adversarial networks." arXiv preprint arXiv:1709.01872 (2017).

[36] Mohebbanaaz, L. V., and Y. Padma Sai. "Classification of Arrhythmia Beats Using Optimized K-Nearest Neighbor Classifier." Intelligent Systems: Proceedings of ICMIB 2020 (2020)

[37] Kumari, Usha, et al. "Feature Extraction and Detection of Obstructive Sleep Apnea from Raw EEG Signal." International Conference on Innovative Computing and Communications. Springer, Singapore, 2020.

[38] Bellemo, Valentina, et al. "Generative adversarial networks (GANs) for retinal fundus image synthesis." Asian Conference on Computer Vision. Springer, Cham, 2018.

[39] Kazeminia, Salome, et al. "GANs for medical image analysis." Artificial Intelligence in Medicine 109 (2020): 101938.

[40] Boudegga, Henda, et al. "Fast and efficient retinal blood vessel segmentation method based on deep learning network." Computerized Medical Imaging and Graphics 90 (2021): 101902.

[41] Pampana, Lakshmi Kala, and Manjula Sri Rayudu. "Retinal Vascular Network Segmentation using a concatenated CNN Architecture." 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON). IEEE, 2021.

A GIS and Fuzzy-based Model for Identification and Analysis of Accident Vulnerable Locations in Urban Traffic Management System: A Case Study of Bhubaneswar

Sarita Mahapatra
Dept. of Comp. Sc. and IT
S'O'A University, Bhubaneswar, India

Krishna Chandra Rath
PG Dept. of Geography
Utkal University, Bhubaneswar, India

Satya Ranjan Das
Dept. of Comp. Sc. and Engg.
S'O'A University, Bhubaneswar, India

***Abstract*—The world has seen road accident and its related societal and economical impact as one of the live, persisting problem in the last 2-3 decades and its prominence has been observed in the developing countries of the Asian subcontinent. With no exception, all major cities in India are facing the various challenges related to road accidents, mostly due to the large population density. Among the major cities in India, Bhubaneswar is a very fast growing city with aim to be the most livable city in Asia in the coming few years. With the city of Bhubaneswar as our study area, we address the issues related to road safety by determining the degree of severity of road accidents. We study the accident related data collected for the last decade using the spatial tools of Geographical Information System (GIS). Then using a GIS-map based analysis and a fuzzy-based model, we have found the spatio-temporal distribution of accident vulnerable locations with their degree of severity. Our experimental results show the accident hot-spots with values of selected contributing parameters such as timing, traffic density, vehicle speed, road intersections.**

***Keywords*—Road traffic management; accident vulnerability; GIS; fuzzy inference; fuzzy rules**

I. Introduction

Every year across the globe, the road accidents are increasing rapidly causing severe injuries and fatalities. As per the road accident data between 2004 to 2013, the accidental death is the 9th major reason of death and going to be 5th major by 2030, unless otherwise proper actions will not be taken [1]. As per [2], the global road accidental death has touched to 1.35 million in the year 2016. Hence, traffic safety is a prominent issue in the sustainable development of urban and semi-urban areas. To ensure traffic safety, road traffic accidents have been considered as one of defining components which contribute to the adverse impact on the economic growth in the developing countries, leading to social and economic concerns. In recent times death cases, severe injuries and property losses in road accidents are the major negative effects of transportation systems. The success of traffic safety programs relies on the analysis of reliable and accurate traffic accident data. Among few notable contributions towards traffic data analysis, in [3], using road accident data of UK between 2005 and 2019 for finding risk and predicting severity of accidental injuries, authors have used data analytics strategies. It can found in [4] that spatial pattern of accident distribution in a specific study area can be analyzed and accident hot spots are identified using GIS based spatial and statistical analysis tools. In [5], authors present the performance result of spatial analysis of accident prone traffic location classification using Multi-criteria Decision Making (MCDM) approach. The study in [6], proposes a detailed spatio-temporal analysis approach combining strategies such as emerging hot spot analysis, spatial autocorrelation analysis and time-space cube analysis to identify the accident hot spots and related traffic features.

This paper analyses the present state of traffic accident information on various arterial and sub-arterial roads in the selected study area that is the city of Bhubaneswar, in the eastern region of India. This paper also discuss the determination of highly accident prone locations using QGIS Software and safety deficient areas on the major roads of the city.

This paper examines the potential use of the Geographic Information Systems (GIS) technology in executing road accident analysis in the study area with the illustration of various point-pattern techniques. Since GIS can further extend the analytical and visualization features, its implementation in urban areas like Bhubaneswar would based more on comprehensive planning and management in data acquisition and integration. A comprehensive digital database on the road structures of Bhubaneswar with an coherent and organized naming conventions is the primary need for the analysis and management of road accidents using GIS development.

In the city map of Bhubaneswar, using GIS-based analysis, locations from the periphery are to be identified in every 500metres; For each location the probability of severity of accidents are to be measured on the basis of five context-specific parameters (described in subsequent discussions); While selecting each parameter in the input interface, a supplementary value is also selected with the parameter. That supplementary value represents the "degree of impact" of each selected parameter. Five possible concepts such as very high, high, normal, low and very low are considered to the context of "degree of impact". To control these fuzzy concepts, a fuzzy controller is used, based on Takagi-Sugeno's fuzzy approach.

Why Fuzzy Classifier: Models based on fuzzy classification are mainly based on processing of many ambiguous, vague, imprecise input information to get some certain inference by

producing various outputs. The input information is processed through a rule-base containing a set of fuzzy if-then-else rules. Depending upon the number and structure of rules in the rule-base, the complexity and quality of fuzzy-classifier is determined.

II. Related Works

P. B. Parmar et al. [7] identified blackspot (i.e. accident-prone locations) on some specific locations of S.P. ring road, Ahmedabad with heatmap plugin and analyzed using QGIS and also suggested remedial measures for the black spot such as to find the places where enforcement steps are needed and number of particular places where sign boards for speed restriction and traffic sign are needed.

J. Choudhary et al. [8] geocoded the accident locations for five years from 2009 to 2013 over the digitized map of the city of Varanasi. The authors also evaluated and analyzed the spatial densities and clustering of accidents using heat-map plugin. The accident hot-spots were isolated with the road stretches using the heat-map using Kernel Density Estimation.

Jerome Ballarta et al. [9] have identified accident hot spots in Katipunan Avenue, Quezon city using standard GIS geoprocessing techniques.

V. Prasannakumar et al. [10] did a comparative investigation on temporal and spatial aspects of road accidents in the city of Thiruvananthapuram. Then using cluster analysis and spatial data statistics, authors have described the temporal and spatial differences of the accident vulnerable locations from the nonvulnerable locations.

Sanjay Kumar Singh and Ashish Misra [11] in their study have analyzed the road accidents in Patna city using annual data in year 1996 to 2000. It provides a overview of road accident scenario in India and deals with existing transport system in Patna.

Deelesh Mandloi and Rajiv Gupta [12] have developed a GIS-based model for predicting the accident vulnerable locations using the road accident related parameters and have also given the remedial steps.

Yen Chen et al. [13] in their paper have discussed on application of geocoding technology for preparing spatial information to related to the traffic accidents and presented a method which accepts potential system reducing accidents as an index to identify the black spots. The association relates the features of road network elements with the black spots, based on the GIS data-storage.

Ela Ertung et al. [14] have carried out a GIS-based analysis of intersection road accidents by creating a database using fatal-injury traffic accident data at intersections in Antalya City Center, Turkey between years 2009 and 2010. They have also determined hot spots for intersection accidents and have conducted statistical evaluations of accidents.

Anik Vega and Dwi Cahyono [15] have used Multiple-Attribute Utility Theory (MAUT) strategy to map dense and accident-vulnerable traffic roads so as to analyze attribute and spatial data. This method helps in finding alternative new road to minimize the traffic density.

Michal Bil et al. [16] in their paper have evaluated and organized clusters of traffic accidents based on their significance. They have also suggested a better strategy for detecting cluster, based on standard Kernel Density Estimation (KDE), suitable to find the most hazardous spots by verifying the importance of the clusters and then ordering the most hazardous spots.

Hao Yu et al. [17] have conducted a comparative investigation of spatial analysis techniques for identifying hot-spots. They collected data from a 622.2-km section on the A1 highway in the UK. From 2001 to 2010, where 7930 crashes were found at the selected highway zones.

Gholam A. Shafabakhsh et al. [18] have conducted spatial analysis of traffic accidents for the city of Mashhad, Iran using GIS methods. Based on the spatial aspect, authors have given a detailed study on various types of road accidents, which is the first attempt in the Mashhad city corporation and also analyzed the accident types using spatial patterns in GIS tool. accidents.

A. Research Contributions

The followings are the list of our contributions in this paper:

1) To study the road accidents data in Odisha from the year 2010 to 2020.
2) To estimate the distribution and incidence of road accidents in the city of Bhubaneswar using GIS.
3) To design a fuzzy rule base with selected set of road accident parameters.
4) To identify the accident vulnerable locations using fuzzy classifier.

III. Study Area

Odisha has registered a sharp rise in road accident fatalities in the five years ending in 2020 against the target reduction in deaths by 50% during same period. A total of 107732 accidents occurred that led to 47884 fatalities. In the road accident statistics of the state of Odisha given in the Fig. 1, the major contributing city is Bhubaneswar, the capital of Odisha. It is seen that the year 2019 has the highest number of fatal accidents, a year which is taken as the starting of the "UN Decade of Action for Road Safety".

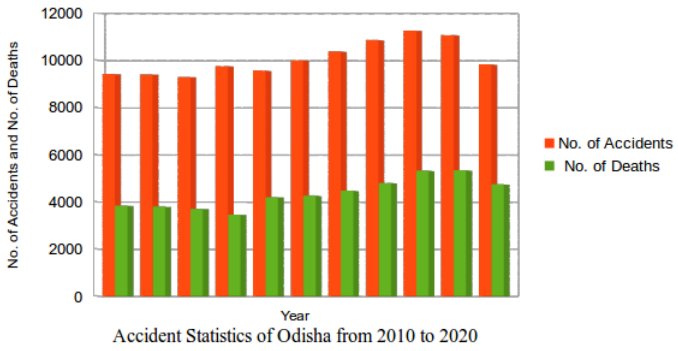


Fig. 1. Year-Wise Accident and Death Statistics of Odisha

Bhubaneswar, the capital of the state of Odisha and one of the popular tourist locations in India, is fast growing city which has organized various international events like International Mega Trade Fair, Men's Hockey Champions Trophy in 2014, Asian Athletics Championships in 2017, Men's Hockey World Cup in 2019, Men's FIH Hockey Junior World Cup in 2021 and also going to organized many more global events including Men's Hockey World Cup in 2023. It is on its way to be India's one of the most well-built cities and selected as a prominent city for the smart-city plan. Hence the crowd density is increasing alarmingly day-by-day leading to more accident cases. The objective of this paper is to evaluate and show hot-spots in Bhubaneswar using information modeling for identifying the statistical locations of accidents using GIS technology and a fuzzy classifier. Spatial-temporal analysis is needed for effective identification of hot spots and can be used to improve the safety of these spots.

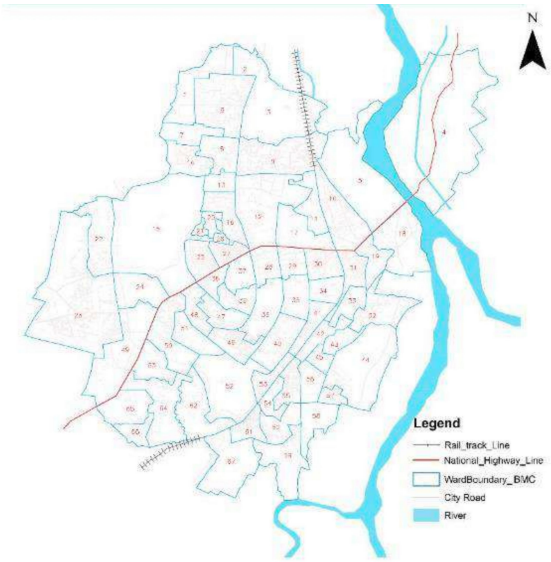


Fig. 2. Road Map of Bhubaneswar

Here Bhubaneswar is considered as the study area which is governed by Bhubaneswar Municipality Corporation. The city of Bhubaneswar sees rapid urbanization with both planned and unplanned way in all sides of the city such as towards Cuttack, Puri and Khurda. In the last three decades, Bhubaneswar has become the business as well as the educational hub of the state along with the best medical facilities available. As a result of which people from all corners of the state and the country as well, are coming to Bhubaneswar for their livelihood. It has a great impact to the population growth, urban environment and most importantly traffic congestion in the city. Nowadays, the traffic issues in the city has become a major concern. In Fig. 2, we have given the road map of Bhubaneswar prepared using QGIS, an open-source GIS software.

Over the years the Ministry of Road Transport and Highways (MoRTH) has been in the process of identifying the accident black spots on the National Highways passing through Bhubaneswar. Most of these spots located on NH-5 (now is NH-16) are mainly because of faulty road designs during ongoing of road projects. Inside the city of Bhubaneswar, Aiginia Chowk and Khandagiri Chowk are found to be two major black spots of NH-16 where 30 and 36 accidental death case occurred respectively between 2017 and 2020. Hence there is an immediate need of micro-level analysis of accident vulnerable locations in Bhubaneswar.

IV. Methodology Used for Identifying Accident Vulnerable Locations in Urban Traffic

Road traffic accidents are now considered as one of the major concerns in India. Authors in [19] mentioned that around 0.4 million accidents are occurring in India almost every year, hence leading to an ever-increasing accident rate. Since accidents are uncertain and unpredictable, hence there is high possibility that this increasing rate of accidents may sustain. Hence there is extreme need of identification of different accident-prone geographical locations and finding the different features related to accidents occurring at these locations which will monitor the different scenarios of road accidents.

In [20], authors suggested that to identify potential measures to prevent road accidents, there is a need for systematic correlation between frequency of accidents and attributes such as traffic information, road-side features, vehicle information, road structure. Lee et al. [21] suggested to use statistical models for analyzing road accidents to determine the correlation among the spatial features and road accident features.

However, [22] found certain demerits of using regular statistical methods for analyzing datasets with large dimensions In addition to problems like sparse data in large dimensional tables, statistical methods can also give some incorrect results mainly due to the assumptions based on the specific models. Hence some techniques based on artificial intelligence and data mining are adopted to handle the large datasets in road accidents.

In [23], authors described some specific data mining applications for road accident analysis, pavement analysis and roughness analysis of road. Authors in [24] narrated the potential data mining strategies like classification, clustering, association rule mining etc. for the effective analysis of road accident data.

As per authors in [25], for effective data analysis the accidents reports available in the police stations are not complete and sufficient for research. At the same time, these basic data can be used and analysed for some specific road segments with the help of statistical approaches, as described in [26], [27].

This paper uses fuzzy logic based classification technique to determine the high-frequency accident spots and subsequent analysis to identify various factors that affect road accidents at these spots. We first identify the parameters for the analysis of the accident vulnerable locations. Then using the fuzzy rule base, we compute the degree of vulnerability of the specific locations by correlating between the characteristics features of these locations and the various attributes in the accident data. Here our main focus is to interpret the results.

A. Proposed Model

The work is basically based on the classification and analysis of traffic data of the city of Bhubaneswar. The

proposed smart traffic management system is going to provide a systematic analysis of data. User has to give the required information like timing (early morning, morning, noon, afternoon, evening, night, late-night), road-condition (i.e. road-friction-density), level-of-road-intersections, existence of crowd-pulling-centers (such as school, bank, shopping mall, etc.) within 200metres periphery and traffic-calming-measures (such as speed-breakers, rumble-strips, bollards), etc. through selecting suitable icons given on the user-interface.

Then the sequence of selected attributes gets associated with values signifying the degree of intensity. The set of user information are then to be stored in a fuzzy rule base which is in the back-end of the system. The fuzzy rule base contains some generalized traffic information and more importantly a set of rules in the form of IF-THEN. In each of the rule the IF-part contains a set of premises which are nothing but the conditions required for finding the degree of accident vulnerability, and the THEN-part contains the response after vulnerability analysis. Hence the response selected by the user is now matched with the premises of the rules in a sequential manner. The rule, for which the inputs matched with all of its premises, will be selected. Then the result part of the selected rule is considered as the response to the user as the result of the vulnerability analysis.

An overview of the proposed system is shown in Fig. 3.

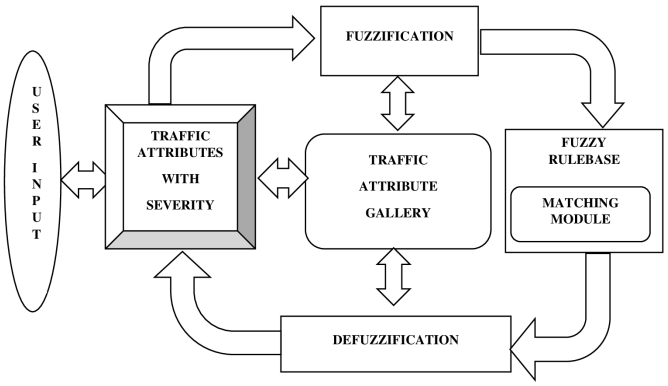


Fig. 3. Framework of Fuzzy-Based Accident-Vulnerability Location Identification

B. Rule-Based Fuzzy Inference System

A rule-base basically consists of the following components:

- a collection of knowledge, in the form of facts
- a set of rules-of-logic, in the form of "if-then", and
- an inference mechanism, which decides what rule is to be used and when.

The knowledge-base in the model consists of these if-then rules which represent conditional statements. Each rule has a premise (P) and a conclusion (C) in the form "if P then C". Based on these rules, the rule-base systems are broadly categorized into two types such as forward-chaining-system which is data-driven and backward-chaining-system which is goal-driven. A forward chaining system is based on first processing the initial data and the using the rules to form the valid conclusions based on the given initial data. A backward chaining system is based on first processing the goals and then searching for rules that can be used for getting the set goals.

C. Rule-based System

The working of the fuzzy controller is described with the following example: The reasoning method in a fuzzy-inference-model is based on the processing of a large set of if-then rules, leading to a knowledge-base. Each rule has a premise, that is the "if" part and a conclusion, that is the "then" part. The knowledge-base in the fuzzy-inference-system (FIS) consists of a large set rules of the form '$if x1 is P1 AND x2 is P2 AND ... xn is Pn then y is C$" where P1, P2, ..., Pn and C are linguistic variables that take fuzzy values from the interval [0,1], and x1, x2, ..., xn are the input parameters and y is the output parameter that is $y = f(x1, x2, ..., xn)$. There are two popular FIS such as Mamdani FIS and Takagi-Sugeno FIS.

In Mamdani-type FIS, the rule-base grows with the increase in parameters in the premise part which leads to difficulty in comprehending the co-relationships between the premises and consequences. In Sugeno-type FIS (or Takagi-Sugeno-Kang method), with fuzzy inputs and a crisp output, It is computationally suitable and efficient to work with adaptive and optimization methods, hence very effective for control problems. Takagi-Sugeno method gives a systematic way to produce fuzzy rules from a given data set. Mamdani-type FIS follows the technique of defuzzification of a fuzzy output where as Sugeno-type FIS follows the computation of crisp output using weighted average. Both the FIS have the same structure in the first two parts such as fuzzifying the inputs and execution of the fuzzy operator. The main difference is in the output membership functions of Sugeno FIS which are either constant or linear.

D. Input Parameters

Based on the data collected for road accidents in the study area, five input parameters are taken which are found to be potential parameters for the prediction of accident vulnerability. Input: $I_1, I_2, I_3, I_4, I_5, a_1, a_2, a_3, a_4, a_5$
where I_1, I_2, I_3, I_4, I_5 represent five parameters which affects vulnerability of locations towards accident, such as traffic density, vehicle speed, presence of crowd-gathering-centers, number of road intersections and timing. The inputs a_1, a_2, a_3, a_4, a_5 are described later in this section.

1) Traffic Density: (I_1) As per the Indian Road Congress (IRC), traffic density can be defined as the average number of vehicles that are present or available in one mile or one kilometer of a road segment, expressed in number of vehicles per kilometer or per mile. Then traffic density (I_1) is computed by:

$$trafficdensity = vehiclecount/segmentlength$$

The survey was conducted two times per hour on selected road portions.

The traffic density (I_1) is computed by the following formula: $I_1 = (1 * C_1 + 2 * C_2 + 3 * C_3 + 4 * C_4)/10$

where C_1, C_2, C_3 and C_4 represent the number of pedestrians, the number of two-wheelers, the number of four and six wheelers, the number of heavy vehicles, respectively.

2) Vehicle Speed: (I_2) Vehicle speed is one important parameter which has taken fuzzy values from [0,1] based on the range of speed.

3) Presence of Crowd Gathering-Centers: (I_3) This parameter is considered as it is including shopping malls, hospitals, educational centers which mainly responsible for crowd gathering and hence leading to accidents and congestions as well.

4) Road Intersections: (I_4) Road intersections can be several types such as 3-ways, 4-ways and multi-ways. Each intersection can affect differently based on the shape, structure, scope, use of channelization and other varieties of traffic-control-devices.

5) Timing: (I_5) Timing of road accidents is one of the most important parameters in the computation of degree of accident vulnerability of locations.

E. Degree of Intensity

We represent the degree of intensity of the accident by a numeric value in between 1 and 5 in the increasing order of the intensities that is from very low to very high. The value of the degree of intensity is based on the values of the input parameters I_1, I_2, I_3, I_4, I_5 and a_1, a_2, a_3, a_4, a_5 represent the degree of intensities of road accidents which are as follows:

- a_j = 5 for I_i = very high(VH), implies highly probable for severe accidents leading to death/fatal cases,
- a_j = 4 for I_i = high(H), implies probable for accidents leading to severe injuries,
- a_j = 3 for I_i = normal(N), implies for probable accidents leading to property loss,
- a_j = 2 for I_i = low(L), implies for probable accidents leading to minor injuries,
- a_j = 1 for I_i = very low(VL), implies for very low probability of accidents,

where $i = 1, ..., 5$ and $j = 1, ..., 5$. In Table I, the range of values for each input parameter is given with respect to their degree of intensities.

V. Experimental Results and Discussions

This section includes the result analysis of the different stages of the input parameter selection and the calculations of the fuzzy output of the rules.

A. Accident Dataset

To test the proposed model, traffic accident data are used which are collected from all districts of Odisha State with special focus on Bhubaneswar. The dataset, provided by the Department of Road Transportation, Odisha and Office of the Police Commissionerate, Bhubaneswar, contains 50,540 traffic accidents that occurred between the year 2011 to 2020. The dataset includes accidents that occurred on national highways such as NH16 and NH203, state highways in Bhubaneswar and its nearby rural areas. The attributes in the dataset include information about the accident timing, vehicle speed, road condition, nearby crowd-gathering centers and traffic density of the road during accident. In the dataset, some attributes are collected from the sensors deployed in the roads. Information, such as actual reason of the accident, light condition, visibility, and road conditions of the accident spot are collected from visibility sensors, traffic surveillance cameras, and tachographs. Hence, due to different modes of data collection, the dataset led to a very complex preprocessing of data. Here we have considered the severity of accident in terms of injury, loss of property, death cases as the dependent variable. Then we have translated these numerical variables into variables with ordinal values. We have considered records of 9640 accidents after preprocessing.

B. Fuzzy Rule-Base

The fuzziness add better and more accurate prediction to the whole system. While selecting value of each parameter as the input, a supplementary value is also selected with the input parameter. That supplementary value represents the "degree of intensity" of each selected parameter. Five possible concepts such as very low, low, normal, severe/high, very severe/very high are considered to the context of "degree of vulnerability". To control these fuzzy concepts, a fuzzy controller is used which is based on Takagi-Sugeno's fuzzy approach. The working of the fuzzy controller is described with the following example:

Input: $I_1, I_2, I_3, I_4, I_5, a_1, a_2, a_3, a_4, a_5$.

where I_1, I_2, I_3, I_4, I_5 represent five parameters, eg. five parameters which affects vulnerability of locations towards accident, eg. traffic density, vehicle speed, presence of crowd-gathering-centres, number of road intersections, timing., and a_1, a_2, a_3, a_4, a_5 represent the degree of intensity of road accidents.

Output: $y_i = f(I_1, I_2, I_3, I_4, I_5) = a_1 * I_1 + a_2 * I_2 + a_3 * I_3 + a_4 * I_4 + a_5 * I_5$

where y_i represents the degree of vulnerability based on the input parameters (I_is) and their degree of intensities (a_js).

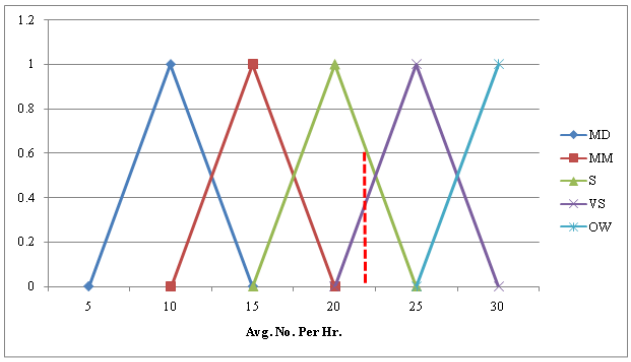


Fig. 4. I_1:Traffic Density

TABLE I. FACTORS OR PARAMETERS WHICH AFFECT VULNERABILITY OF LOCATIONS

Parameters	Very High(0.9-1.0)	High(0.7-0.8)	Normal(0.5-0.6)	Low(0.3-0.4)	Very Low(0.1-0.2)
traffic density	0.9-1.0	0.7-0.8	0.5-0.6	0.3-0.4	0.1-0.2
vehicle speed	> 80	60-80	40-60	20-40	0-20
road-intersections	> 6	6	5	4	3
crowd-gathering-centers	> 5	4	3	2	1
timing	15:00-21:00	9:00-15:00	6:00-9:00	3:00-6:00	00:00-3:00

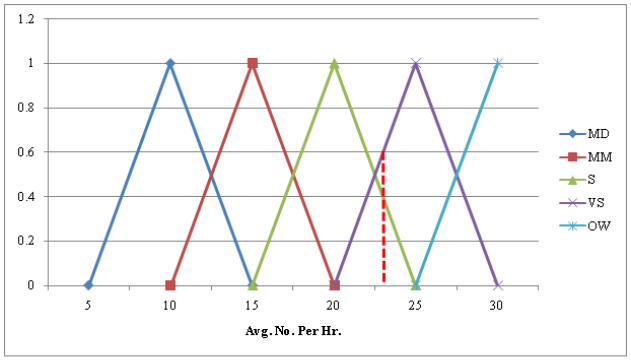


Fig. 5. I_2:Vehicle Speed

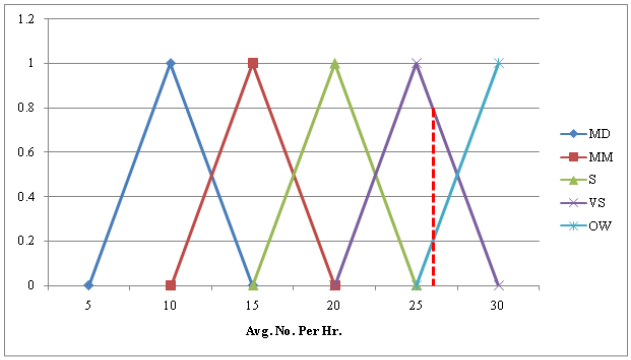


Fig. 7. I_4:Number of Road Intersections

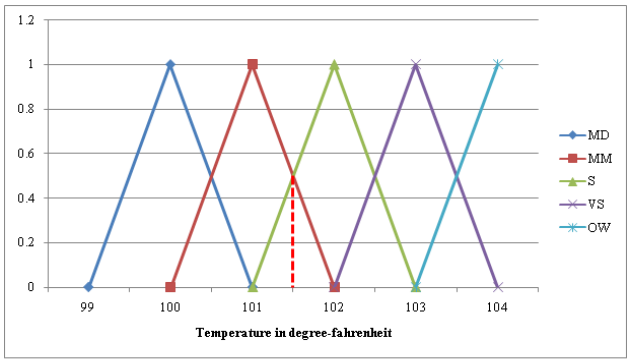


Fig. 6. I_3:Presence of Crowd-Gathering-Centres

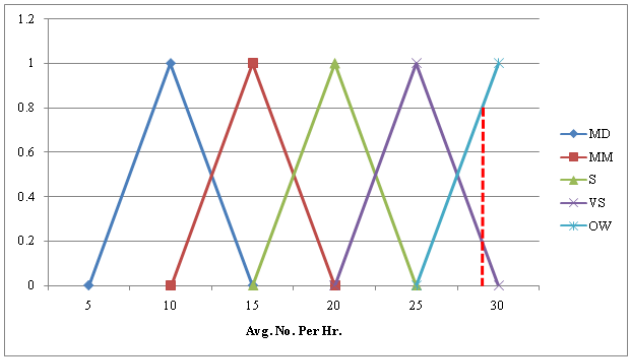


Fig. 8. $I_5$1:Accident Timing

We have taken an example scenario, which is as follows. Let $I_1 = 0.6$ per kilometer, from the Fig. 4 using triangle formula we get:

$$\mu_H(0.6) = 0.6$$
$$\mu_{VH}(0.6) = 0.4$$

Let $I_{82} = 82$ kilometers per hour, from the Fig. 5 using triangle formula we get:

$$\mu_H(82) = 0.7$$
$$\mu_{VH}(82) = 0.8$$

Let $I_3 = 3$ number of crowd-gathering-centre, from the Fig. 6 using triangle formula we get:

$$\mu_N(3) = 0.5$$
$$\mu_H(3) = 0.5$$

Let $I_4 = 4$ number of road intersections, from the Fig. 7 using triangle formula we get:

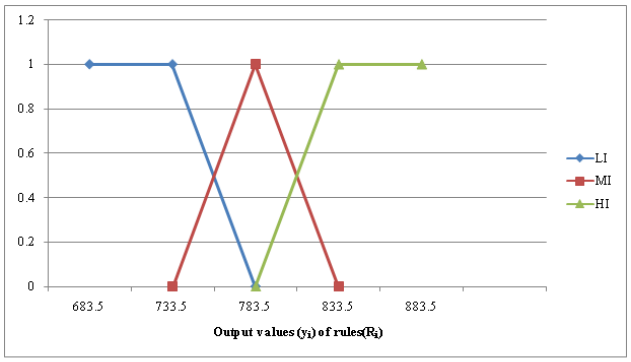


Fig. 9. Fuzzy Output of Rules

$\mu_H(4) = 0.8$
$\mu_{VH}(4) = 0.2$

Let $I_5 = 20$, from the Fig. 8 using triangle formula we get:

$\mu_{VH}(20) = 0.9$
$\mu_N(20) = 0.6$

From the above analysis, we get 32 numbers of rules. For each rule R_i, the corresponding y_i and ω_i (represents weight of rule R_i) value is computed as below:

Rule R_1: (I_1 is H)(I_2 is S)(I_3 is S)(I_4 is H)(I_5 is H)
$\omega_1 = \mu_H * \mu_S * \mu_S * \mu_H * \mu_H = 0.6 * 0.4 * 0.5 * 0.8 * 0.2 = 0.0192$
$y_1 = 4 * I_1 + 3 * I_2 + 3 * I_3 + 4 * I_4 + 4 * I_5 = 681.5$

Rule R_2: (I_1 is H)(I_2 is S)(I_3 is S)(I_4 is H)(I_5 is VH)
$\omega_2 = \mu_H * \mu_S * \mu_S * \mu_H * \mu_{VH} = 0.6 * 0.4 * 0.5 * 0.8 * 0.8 = 0.0768$
$y_2 = 4 * I_1 + 3 * I_2 + 3 * I_3 + 4 * I_4 + 5 * I_5 = 710.5$

Rule R_3: (I_1 is H)(I_2 is S)(I_3 is S)(I_4 is VH)(I_5 is H)
$\omega_3 = \mu_H * \mu_S * \mu_S * \mu_{VH} * \mu_H = 0.6 * 0.4 * 0.5 * 0.2 * 0.2 = 0.0048$
$y_3 = 4 * I_1 + 3 * I_2 + 3 * I_3 + 5 * I_4 + 4 * I_5 = 707.5$

Rule R_4: (I_1 is H)(I_2 is S)(I_3 is S)(I_4 is VH)(I_5 is VH)
$\omega_4 = \mu_H * \mu_S * \mu_S * \mu_{VH} * \mu_{VH} = 0.6 * 0.4 * 0.5 * 0.2 * 0.8 = 0.0192$
$y_4 = 4 * I_1 + 3 * I_2 + 3 * I_3 + 5 * I_4 + 5 * I_5 = 736.5$

Rule R_5: (I_1 is H)(I_2 is S)(I_3 is H)(I_4 is H)(I_5 is H)
$\omega_5 = \mu_H * \mu_S * \mu_H * \mu_H * \mu_H = 0.6 * 0.4 * 0.5 * 0.8 * 0.2 = 0.0192$
$y_5 = 4 * I_1 + 3 * I_2 + 4 * I_3 + 4 * I_4 + 4 * I_5 = 783.0$

Rule R_6: (I_1 is H)(I_2 is S)(I_3 is H)(I_4 is H)(I_5 is VH)
$\omega_6 = \mu_H * \mu_S * \mu_H * \mu_H * \mu_{VH} = 0.6 * 0.4 * 0.5 * 0.8 * 0.8 = 0.0768$
$y_6 = 4 * I_1 + 3 * I_2 + 4 * I_3 + 4 * I_4 + 5 * I_5 = 812.0$

Rule R_7: (I_1 is H)(I_2 is S)(I_3 is H)(I_4 is VH)(I_5 is H)
$\omega_7 = \mu_H * \mu_S * \mu_H * \mu_{VH} * \mu_H = 0.6 * 0.4 * 0.5 * 0.2 * 0.2 = 0.0048$
$y_7 = 4 * I_1 + 3 * I_2 + 4 * I_3 + 5 * I_4 + 4 * I_5 = 809.0$

Rule R_8: (I_1 is H)(I_2 is S)(I_3 is H)(I_4 is VH)(I_5 is VH)
$\omega_8 = \mu_H * \mu_S * \mu_H * \mu_{VH} * \mu_{VH} = 0.6 * 0.4 * 0.5 * 0.2 * 0.8 = 0.0192$
$y_8 = 4 * I_1 + 3 * I_2 + 4 * I_3 + 5 * I_4 + 5 * I_5 = 838.0$

Rule R_9: (I_1 is H)(I_2 is H)(I_3 is S)(I_4 is H)(I_5 is H)
$\omega_9 = \mu_H * \mu_H * \mu_S * \mu_H * \mu_H = 0.6 * 0.6 * 0.5 * 0.8 * 0.2 = 0.0288$
$y_9 = 4 * I_1 + 4 * I_2 + 3 * I_3 + 4 * I_4 + 4 * I_5 = 704.5$

Rule R_{10}: (I_1 is H)(I_2 is H)(I_3 is S)(I_4 is H)(I_5 is VH)
$\omega_{10} = \mu_H * \mu_H * \mu_S * \mu_H * \mu_{VH} = 0.6 * 0.6 * 0.5 * 0.8 * 0.8 = 0.1152$
$y_{10} = 4 * I_1 + 4 * I_2 + 3 * I_3 + 4 * I_4 + 5 * I_5 = 733.5$

Rule R_{11}: (I_1 is H)(I_2 is H)(I_3 is S)(I_4 is VH)(I_5 is H)
$\omega_{11} = \mu_H * \mu_H * \mu_S * \mu_{VH} * \mu_H = 0.6 * 0.6 * 0.5 * 0.2 * 0.2 = 0.0072$
$y_{11} = 4 * I_1 + 4 * I_2 + 3 * I_3 + 5 * I_4 + 4 * I_5 = 730.5$

Rule R_{12}: (I_1 is H)(I_2 is H)(I_3 is S)(I_4 is VH)(I_5 is VH)
$\omega_{12} = \mu_H * \mu_H * \mu_S * \mu_{VH} * \mu_{VH} = 0.6 * 0.6 * 0.5 * 0.2 * 0.8 = 0.0288$
$y_{12} = 4 * I_1 + 4 * I_2 + 3 * I_3 + 5 * I_4 + 5 * I_5 = 759.5$

Rule R_{13}: (I_1 is H)(I_2 is H)(I_3 is H)(I_4 is H)(I_5 is H)
$\omega_{13} = \mu_H * \mu_H * \mu_H * \mu_H * \mu_H = 0.6 * 0.6 * 0.5 * 0.8 * 0.2 = 0.0288$
$y_{13} = 4 * I_1 + 4 * I_2 + 4 * I_3 + 4 * I_4 + 4 * I_5 = 806.0$

Rule R_{14}: (I_1 is H)(I_2 is H)(I_3 is H)(I_4 is H)(I_5 is VH)
$\omega_{14} = \mu_H * \mu_H * \mu_H * \mu_H * \mu_{VH} = 0.6 * 0.6 * 0.5 * 0.8 * 0.8 = 0.1152$
$y_{14} = 4 * I_1 + 4 * I_2 + 4 * I_3 + 4 * I_4 + 5 * I_5 = 835.0$

Rule R_{15}: (I_1 is H)(I_2 is H)(I_3 is H)(I_4 is VH)(I_5 is H)
$\omega_{15} = \mu_H * \mu_H * \mu_H * \mu_{VH} * \mu_H = 0.6 * 0.6 * 0.5 * 0.2 * 0.2 = 0.0072$
$y_{15} = 4 * I_1 + 4 * I_2 + 4 * I_3 + 5 * I_4 + 4 * I_5 = 832.0$

Rule R_{16}: (I_1 is H)(I_2 is H)(I_3 is H)(I_4 is VH)(I_5 is VH)
$\omega_{16} = \mu_H * \mu_H * \mu_H * \mu_{VH} * \mu_{VH} = 0.6 * 0.6 * 0.5 * 0.2 * 0.8 = 0.0288$
$y_{16} = 4 * I_1 + 4 * I_2 + 4 * I_3 + 5 * I_4 + 5 * I_5 = 861.0$

Rule R_{17}: (I_1 is VH)(I_2 is S)(I_3 is S)(I_4 is H)(I_5 is H)
$\omega_{17} = \mu_{VH} * \mu_S * \mu_S * \mu_H * \mu_H = 0.4 * 0.4 * 0.5 * 0.8 * 0.2 = 0.0128$
$y_{17} = 5 * I_1 + 3 * I_2 + 3 * I_3 + 4 * I_4 + 4 * I_5 = 703.5$

Rule R_{18}: (I_1 is VH)(I_2 is S)(I_3 is S)(I_4 is H)(I_5 is

VH)
$\omega_{18} = \mu_{VH} * \mu_S * \mu_S * \mu_H * \mu_{VH} = 0.4 * 0.4 * 0.5 * 0.8 * 0.8 = 0.0512$
$y_{18} = 5 * I_1 + 3 * I_2 + 3 * I_3 + 4 * I_4 + 5 * I_5 = 732.5$

Rule R_{19}: (I_1 is VH)(I_2 is S)(I_3 is S)(I_4 is VH)(I_5 is H)
$\omega_{19} = \mu_{VH} * \mu_S * \mu_S * \mu_{VH} * \mu_H = 0.4 * 0.4 * 0.5 * 0.2 * 0.2 = 0.0032$
$y_{19} = 5 * I_1 + 3 * I_2 + 3 * I_3 + 5 * I_4 + 4 * I_5 = 729.5$

Rule R_{20}: (I_1 is VH)(I_2 is S)(I_3 is S)(I_4 is VH)(I_5 is VH)
$\omega_{20} = \mu_{VH} * \mu_S * \mu_S * \mu_{VH} * \mu_{VH} = 0.4 * 0.4 * 0.5 * 0.2 * 0.8 = 0.0128$
$y_{20} = 5 * I_1 + 3 * I_2 + 3 * I_3 + 5 * I_4 + 5 * I_5 = 758.5$

Rule R_{21}: (I_1 is VH)(I_2 is S)(I_3 is H)(I_4 is H)(I_5 is H)
$\omega_{21} = \mu_{VH} * \mu_S * \mu_H * \mu_H * \mu_H = 0.4 * 0.4 * 0.5 * 0.8 * 0.2 = 0.0128$
$y_{21} = 5 * I_1 + 3 * I_2 + 4 * I_3 + 4 * I_4 + 4 * I_5 = 805.0$

Rule R_{22}: (I_1 is VH)(I_2 is S)(I_3 is H)(I_4 is H)(I_5 is VH)
$\omega_{22} = \mu_{VH} * \mu_S * \mu_H * \mu_H * \mu_{VH} = 0.4 * 0.4 * 0.5 * 0.8 * 0.8 = 0.0512$
$y_{22} = 5 * I_1 + 3 * I_2 + 4 * I_3 + 4 * I_4 + 5 * I_5 = 834.0$

Rule R_{23}: (I_1 is VH)(I_2 is S)(I_3 is H)(I_4 is VH)(I_5 is H)
$\omega_{23} = \mu_{VH} * \mu_S * \mu_H * \mu_{VH} * \mu_H = 0.4 * 0.4 * 0.5 * 0.2 * 0.2 = 0.0032$
$y_{23} = 5 * I_1 + 3 * I_2 + 4 * I_3 + 5 * I_4 + 4 * I_5 = 831.0$

Rule R_{24}: (I_1 is VH)(I_2 is S)(I_3 is H)(I_4 is VH)(I_5 is VH)
$\omega_{24} = \mu_{VH} * \mu_S * \mu_H * \mu_{VH} * \mu_{VH} = 0.4 * 0.4 * 0.5 * 0.2 * 0.8 = 0.0128$
$y_{24} = 5 * I_1 + 3 * I_2 + 4 * I_3 + 5 * I_4 + 5 * I_5 = 860.0$

Rule R_{25}: (I_1 is VH)(I_2 is H)(I_3 is S)(I_4 is H)(I_5 is H)
$\omega_{25} = \mu_{VH} * \mu_H * \mu_S * \mu_H * \mu_H = 0.4 * 0.6 * 0.5 * 0.8 * 0.2 = 0.0192$
$y_{25} = 5 * I_1 + 4 * I_2 + 3 * I_3 + 4 * I_4 + 4 * I_5 = 726.5$

Rule R_{26}: (I_1 is VH)(I_2 is H)(I_3 is S)(I_4 is H)(I_5 is VH)
$\omega_{26} = \mu_{VH} * \mu_H * \mu_S * \mu_H * \mu_{VH} = 0.4 * 0.6 * 0.5 * 0.8 * 0.8 = 0.0768$
$y_{26} = 5 * I_1 + 4 * I_2 + 3 * I_3 + 4 * I_4 + 5 * I_5 = 755.5$

Rule R_{27}: (I_1 is VH)(I_2 is H)(I_3 is S)(I_4 is VH)(I_5 is H)
$\omega_{27} = \mu_{VH} * \mu_H * \mu_S * \mu_{VH} * \mu_H = 0.4 * 0.6 * 0.5 * 0.2 * 0.2 = 0.0048$
$y_{27} = 5 * I_1 + 4 * I_2 + 3 * I_3 + 5 * I_4 + 4 * I_5 = 752.5$

Rule R_{28}: (I_1 is VH)(I_2 is H)(I_3 is S)(I_4 is VH)(I_5 is VH)
$\omega_{28} = \mu_{VH} * \mu_H * \mu_S * \mu_{VH} * \mu_{VH} = 0.4 * 0.6 * 0.5 * 0.2 * 0.8 = 0.0192$
$y_{28} = 5 * I_1 + 4 * I_2 + 3 * I_3 + 5 * I_4 + 5 * I_5 = 781.5$

Rule R_{29}: (I_1 is VH)(I_2 is H)(I_3 is H)(I_4 is H)(I_5 is H)
$\omega_{29} = \mu_{VH} * \mu_H * \mu_H * \mu_H * \mu_H = 0.4 * 0.6 * 0.5 * 0.8 * 0.2 = 0.0192$
$y_{29} = 5 * I_1 + 4 * I_2 + 4 * I_3 + 4 * I_4 + 4 * I_5 = 828.0$

Rule R_{30}: (I_1 is VH)(I_2 is H)(I_3 is H)(I_4 is H)(I_5 is VH)
$\omega_{30} = \mu_{VH} * \mu_H * \mu_H * \mu_H * \mu_{VH} = 0.4 * 0.6 * 0.5 * 0.8 * 0.8 = 0.0768$
$y_{30} = 5 * I_1 + 4 * I_2 + 4 * I_3 + 4 * I_4 + 5 * I_5 = 857.0$

Rule R_{31}: (I_1 is VH)(I_2 is H)(I_3 is H)(I_4 is VH)(I_5 is H)
$\omega_{31} = \mu_{VH} * \mu_H * \mu_H * \mu_{VH} * \mu_H = 0.4 * 0.6 * 0.5 * 0.2 * 0.2 = 0.0048$
$y_{31} = 5 * I_1 + 4 * I_2 + 4 * I_3 + 5 * I_4 + 4 * I_5 = 854.0$

Rule R_{32}: (I_1 is VH)(I_2 is H)(I_3 is H)(I_4 is VH)(I_5 is VH)
$\omega_{32} = \mu_{VH} * \mu_H * \mu_H * \mu_{VH} * \mu_{VH} = 0.4 * 0.6 * 0.5 * 0.2 * 0.8 = 0.0192$
$y_{32} = 5 * I_1 + 4 * I_2 + 4 * I_3 + 5 * I_4 + 5 * I_5 = 883.0$

Now the combined action of all the rules can be obtained as follows:

$$y = \sum_{i=1}^{32} \omega_i y_i / \sum_{j=1}^{32} \omega_j = 783.25/1.0 = 783.25$$

Hence from the above computations we infer that: the rules (R_is) with output values (y_is) greater than the value of y are cases with high vulnerability and hence alertness is to be given in the output interface. The output of the fuzzy-controller, that is the degree of vulnerability is defined using fuzzy values as, given in the Fig. 9:

$$b_i(y_i) = \begin{cases} 1(LV) & \text{for} \quad y_i < 733.5 \\ 2(MV) & \text{for} \quad 733.5 \leq y_i \leq 833.5 \\ 3(HV) & \text{for} \quad y_i > 833.5 \end{cases}$$

Here LV, MV and HV represent low vulnerable, medium vulnerable and high vulnerable respectively. Hence, sequence of control is as follows:

parameters selected ⇒ for each input parameter, the degree of intensity selected ⇒ corresponding fuzzy ruled fired ⇒ with the output of the fuzzy rule, the degree of vulnerability notified.

Based on the values of the selected input parameters, the degree of vulnerabilities are computed for some specific locations in our study area. It is found that in addition to Khandagiri and Aiginia square, the other black spots identified includes Patrapada, Palasuni and Hanspal within the limits of our study area Bhubaneswar.

VI. Conclusion

Fuzzy Inference Systems have been reliable strategies for the analysis of accident data sets taken from various countries of the globe. In several attempts, different authors have used the Mamdani and Sugeno FIS for finding the causes of road accident severity. In this paper, we have taken an integrated

approach of GIS data modeling with fuzzy inference mechanism to analyze and identify the accident vulnerable locations withe their degree of vulnerability in the study area.

The present study have used fuzzy-based technique on different sections of accident locations. The rules generated for each section expressed the various reasons associated with road accidents in the specific locations. Each section may contain some similar rules, but they have different values for each group. The dataset for road accident and its analysis using fuzzy-based method shows that this method can be applied on other accident data having larger number of attributes to find more number of parameters linked with road accidents. It is observed that this fuzzy-based method has adequately found reasonable information from the given data set, with the outcomes produced at very general level because of some missing information such as the victim information, road surface condition, weather related information. The data with higher number of attributes can extract additional information using the present strategy.

REFERENCES

[1] A. A. Hyder, N. Paichadze, T. Toroyan, and M. M. Peden, —Monitoring the Decade of Action for Global Road Safety 2011–2020: An update, Glob. Public Health,vol. 12, no. 12, pp. 1492–1505, 2017.

[2] Organization, W.H., "Global Status Report on Road Safety 2018," WHO: Geneva, Switzerland. 2019

[3] Mohamed K Nour, Atif Naseer, Basem Alkazemi, Muhammad Abid Jamil. "Road Traffic Accidents Injury Data Analytics", (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 11, No. 12, 2020.

[4] Wesam Alkhadour, Jamal Zraqou, Adnan Al-Helali, Sajeda Al-Ghananeem. "Traffic Accidents Detection using Geographic Information Systems (GIS)Spatial Correlation of Traffic Accidents in the City of Amman, Jordan", (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 12, No.4, 2021.

[5] Anik Vega Vitianingsih, Zahriah Othman, Safiza Suhana Kamal Baharin, Aji Suraji. "Empirical Study of a Spatial Analysis for Prone Road Traffic Accident Classification based on MCDM Method", (IJACSA) International Journal of Advanced Computer Science and ApplicationsVol. 13, No. 5, 2022.

[6] Cheng Z, Zu Z, Lu J (2018) Traffic crash evolution characteristic analysis and spatiotemporal hotspot identification of urban road intersections. Sustainability 2019(11):160. https://doi.org/10.3390/su11010160

[7] Parmar, P. B.(2018). Black Spot Analysis Using QGIS for S.P. Ring Road, Ahmedabad (Ch.: 00.00Km To Ch.:76.30Km). International Research Journal of Engineering and Technology.

[8] Choudhary, Jayvant & Ohri, Anurag & Kumar, Brind. (2015). Identification of Road Accidents Hot Spots in Varanasi using QGIS. Proceedings of National Conference on Open-Source GIS: Opportunities and Challenges; Department of Civil Engineering,IIT(BHU), Varanasi; October 9-10, 2015. ISBN:978-81-931-2500-7

[9] Villanueva, Cecilia May & Doroy, Nelson & Ballarta, Jerome & Padao, Ishtar. (2015). Accident Hotspot Mapping of Katipunan Avenue, Quezon City. EASTS 2015; Cebu City, Philippines.

[10] Prasannakumar, V. & H., Vijith & Charutha, R. & Geetha, N.. (2011). Spatio-Temporal Clustering of Road Accidents: GIS Based Analysis and Assessment. Procedia - Social and Behavioral Sciences. 21. 317-325. 10.1016/j.sbspro.2011.07.020.

[11] Singh, Sanjay Kumar and Misra, Ashish, Road Accident Analysis: A Case Study of Patna City. Urban Transport Journal, Vol. 2, No. 2, pp. 60-75, 2004, Available at SSRN: https://ssrn.com/abstract=570441

[12] Mandloi, Deelesh & Gupta, Rajiv. (2003). Evaluation of accident black spots on roads using Geographical Information Systems (GIS). Map India Conference 2003.

[13] Chen, Yan & Wu, Hangbin & Liu, Chun & Sun, Weiwei. (2011). Identification of black spot on traffic accidents and its spatial association analysis based on geographic information system. Proceedings - 2011 7th International Conference on Natural Computation, ICNC 2011. 1. 143-150. 10.1109/ICNC.2011.6021904.

[14] Ela Ertunc, Tayfun Cay, Omer Mutluoglu. "Analysis of Road Traffic Accident in Antalya Province (Turkey) Using Geographical Information Systems". SUJEST, v.4, n.4, 2016; ISSN: 2147-9364.

[15] A. V. Vitianingsih and D. Cahyono, "Geographical Information System for mapping accident-prone roads and development of new road using Multi-Attribute Utility method," 2016 2nd International Conference on Science and Technology-Computer (ICST), 2016, pp. 66-70, doi: 10.1109/ICSTC.2016.7877349.

[16] Bíl, Michal & Andrášik, Richard & Janoška, Zbyněk. (2013). Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. Accident; analysis and prevention. 55C. 265-273. 10.1016/j.aap.2013.03.003.

[17] Yu H, Liu P, Chen J, Wang H. Comparative analysis of the spatial analysis methods for hotspot identification. Accid Anal Prev. 2014 May;66:80-8. doi: 10.1016/j.aap.2014.01.017. Epub 2014 Jan 29. PMID: 24530515.

[18] Gholam Ali Shafabakhsh, Afshin Famili, Mohammad Sadegh Bahadori, GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran, Journal of Traffic and Transportation Engineering (English Edition), Volume 4, Issue 3, 2017, Pages 290-299, ISSN 2095-7564, https://doi.org/10.1016/j.jtte.2017.05.005.

[19] MORTH (2014) Road Accidents in India 2013. New Delhi: Ministry of Road Transport and Highways Transport Research Wing, Government of India, August 2014. http://morth.nic.in/ showfile.asp?lid=1465. Accessed 20 May 2015. showfile.asp?lid=1465. Accessed 20 May 2015.

[20] Kononov J, Janson BN (2002) Diagnostic methodology for the detection of safety problems at intersections. Transp Res Rec. doi:10.3141/1784-07

[21] Lee C, Saccomanno F, Hellinga B (2002) Analysis of crash precursors on instrumented freeways. Transp Res Rec. doi:10.3141/1784-01

[22] Chen W, Jovanis P (2000) Method for identifying factors contributing to driver-injury severity in traffic crashes. Transp Res Rec. doi:10.3141/1717-01

[23] Barai S (2003) Data mining application in transportation engineering. Transport 18:216–223. doi:10.1080/16483840.2003.10414100

[24] Tan PN, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson Addison-Wesley, Boston

[25] Ponnaluri RV (2012) Road traffic crashes and risk groups in India: analysis, interpretations, and prevention strategies. IATSS Res 35:104–110. doi:10.1016/j.iatssr.2011.09.002

[26] Kumar CN, Parida M, Jain SS (2013) Poisson family regression techniques for prediction of crash counts using Bayesian inference. Proc Soc Behav Sci 104:982–991. doi:10.1016/j.sbspro.2013.11.193

[27] Parida M, Jain SS, Kumar CN (2012) Road traffic crash prediction on national highways. Indian Highway Road Congress 40:93–103

Novel Deep Learning Technique to Improve Resolution of Low-Quality Finger Print Image for Bigdata Applications

Lisha P P
Research Scholar
Govt. Model Engineering College
Ernakulam, Kerala, India 682021

Jayasree V K
Professor
Dept. of Electronics and Communication
Govt. Model Engineering College
Ernakulam, Kerala, India 682021

***Abstract*—High-resolution images are highly in demand when they are utilized for different analysis purposes and obviously due to their quality aesthetic visual impact. The objective of image super-resolution (SR) is to reconstruct a high-resolution (HR) image from a low-resolution (LR)image. Storing, transferring and processing of high-resolution (HR) images have got many practical issues in big data domain. In the case of finger print images, the data is huge because of the huge number of populations. So instead of transferring or storing these finger print images in its original form (HR images), it cost very low if we choose its low-resolution form. By using sampling technique, we can easily generate LR images, but the main problem is to regenerate HR image from these LR images. So, this paper addresses this problem, a novel method for enhancing resolution of low-resolution fingerprint images of size 50 x 50 to a high-resolution image of size 400 x 400 using convolutional neural network (CNN) architecture followed by sub pixel convolution operation for up sampling with no loss of promising features available in low-resolution image has been proposed. The proposed model contains five convolutional layers, each of which has an appropriate number of filter channels, activation functions, and optimization functions. The proposed model was trained using three publicly accessible fingerprint datasets FVC 2004 DB1, DB2, and DB3 after being validation and testing were done using 10 percent of these fingerprint data sets. In terms of performance measures like Peak Signal to Noise Ratio (PSNR), Mean Squared Error (MSE), Structural Similarity Index (SSIM) and loss functions, the quantitative and qualitative results show that the proposed model greatly outperformed existing state-of-the-art techniques like Enhanced deep residual network (EDSR), wide activation for image and video SR (WDSR), Generative adversarial network(GAN) based models and Auto-encoder-based models.**

Keywords*—Single image super-resolution; convolution neural network; biometric; fingerprint images

I. Introduction

A biometric system is an effective tool used for personal identification in the fields of healthcare, insurance, forensics, security systems etc. In the modern computer age, among all other biometric systems, the fingerprint is one of the most widely used biometric systems. It is utilised for both personal identification and verification [1].

Now a days, in a fingerprint's rich details, such as pores and scars, can be captured using an optical fingerprint sensor at a resolution of 2000 dpi. Large scale bio metric systems are employed for personal identifications in nationally accepted identity cards and also in mobile based payment applications. For example, in India, nation-wide identity card like Aadhar, where each of the 1.3 billion citizens will have two iris pictures and all 10 of their fingerprints stored in huge database [2]. In big data system like this, it is a difficult task to meet hardware and software needed for this enormous storage and analysis for personal identification. It is challenging to compare an unknown fingerprint with this vast amount of data in personal identification. The potential of deep learning in big data allows for the analysis of very large and complex data, including images, videos or text, in the field of healthcare. Big data has security and privacy issues since acessability and process used to store, manipulate and retain data has increased [3]. In this scenario, a method for personal identification and analysis using fingerprint image with minimum computations and storage is highly in demand. Super-resolution techniques for enhancing low-resolution images in to high-resolution images with no loss of promising features used for identification is suitable in this situation, Here, we need to store low resolution images of size 50 x 50, but for fingerprint image analysis or verification, it is enhanced in to 8 times than input image so it will drastically reduce complexity involved in both hardware and software for storage as well as analysing huge volume of data.

While dealing with natural scenes and situations, the availability of high-resolution images is not always effortless. The major hurdles for the same are noise, blur, camera limitations, and limitations in acquisitions. The domains including medical diagnosis, digital surveillance, remote sensing and forensics analysis always require high-resolution images [4]. In forensics, biometric has prominent role and mostly fingerprint has vital role. For all those applications where personal identification and verification employed using fingerprint image, require high quality image and also image analysis of the same is to be performed with less space and computational complexity.

Super-Resolution (SR) is the means of reconstructing a high-quality image utilizing one or more low-quality image(s). Super-resolution techniques are divided into two categories: single image super-resolution (SISR) and multi-image super-resolution (MISR). SISR reconstructs the SR image from a low-resolution input image. SR models in the SISR and MISR categories have been constructed using either classical or deep learning methods. In the classical approach, Andrew Gilman et

al. [5] observed multiple algorithms for each super-resolution category and they are given as interpolation-based, learning-based, and reconstruction-based. Bi-linear, bi-cubic, and cubic spline are the most often used interpolation algorithms. These approaches used weighted average of neighbouring LR pixels to estimate unknown HR pixels. Ledig et al. [6] observed that interpolation-based approaches are quick and easy, but they muddy the image's details and make it difficult to establish the image's precision hence blurring of features and edges in a sample image is caused. Using an external image data-set, learning-based algorithms build a link between the LR and HR image. Nasrollahi, K., and Moeslund, T.B [7] observed that in reconstruction-based methods, the details of the HR image such as edge prior, gradient prior are recovered using some prior knowledge. In classical methods, images cannot be magnified beyond the image sample resolution without losing image quality. In recent years, deep learning-based super-resolution models have superior performance over classical methods in all applications in which image analysis is required.

In this paper a novel method for generating high-resolution fingerprint image from a low-resolution fingerprint image is proposed in which resolution enhanced eight times than input image using convolution neural network architecture and sub pixel convolution operation.

The paper is organized in such a way that the related work for SISR approaches is offered in Section II, the framework for the proposed model is explained in Section III, Experiment part is described in Section IV, results and evaluation metrics are reviewed in Section V, and the conclusion is presented in Section VI.

II. Related Work

Major approaches for super resolution of fingerprint images includes classical image processing techniques, Residual networks-based SR models, Auto-encoder based SR, Convolutional neural network-based SR and Generative adversarial network-based SR models as described below.

A. Classical Image Processing Techniques for SR

Ganchimeg G and Leopold H developed a model for fingerprint enhancement based on classical image processing techniques [8]. They employed filtering methods for noise removal followed by edge detection methods and thinning process for enhancement, but their result gets blurred for higher magnification orders. Nouf Saeed and Alotaibi [9] applied Gabor filter for denoising and after that deep boltzmann method is applied for ridge enhancement. Dinca Lazarescu Andreea-Monica et al.[10] applied convolutional layers for feature extraction and also for mapping low-resolution fingerprint image in to high-resolution images, but in their method, initially low-resolution image is enhanced using Laplacian filters and thereafter convolutional operations are employed. Major limitation of classical image processing techniques is artifacts or blurring occurs for large magnification orders.

B. Residual Network based SR Models

Zhenzhen Yang et al. [11] applied residual network for enhancing personal identifying features pores and ridges available in low resolution images. In their model, they applied 32 residual blocks in which each block contains 3 convolutional layers and relu activation functions. Finally resultant features extracted using residual blocks are combined with pixel shuffling blocks and regenerated output image. Since model is complex, it is computationally intensive. Seonjae K [12] employed two CNN (Convolutional neural network) based network for feature extraction and enhancement. First network employed for feature extraction of low-resolution images using dense layer with local skip connections. After that input image upscaled using interpolation technique and passed in to second network, which composed of many dense layers with local and global skip connections. Finally output from first and second network concatenated to regenerate enhanced image. Here, since second network performs feature extraction on interpolated image, computational requirement is considerably more and also dense network processing on features of interpolated image rather than original image. Yongliang Zhang et al. [13] employed convolutional residual network for fingerprint liveness detection.

C. Autoencoder based SR Models

Sandoval Veríssimo de Sousa Neto et al. [14] applied deep convolutional auto-encoder for feature extraction and applied Gabor filtering and Gaussian filtering for enhancement. But major limitation of auto-encoders is while encoding input image in to latent vectors, information loss may occur. Sergio Saponara et al. [15] applied convolutional auto-encoders for fingerprint image enhancement. In their model they employed convolutional layers for feature extraction, max-pooling layer used for down sampling and up sampling by deconvolution layer. Major drawback of auto-encoder based models are, while encoding input image, all promising features may not be represented in latent form hence during decoding and regenerating phase information loss takes place.

D. Convolutional Neural Network (CNN) based SR Models

Ajnas muhammed and Alwyn roshan [16] employed deep convolutional network (20 layers) for image enhancement. Since their network is deeper, it is computationally expensive. Ayushi Tamrakar and Neetesh Gupta [17] proposed SR model based on convolutional neural network (CNN) and long short-term memory (LSTM). In their model, they first applied CNN for both feature extraction and enhancement, thereafter LSTM applied on this feature map to classify images based on ridges in output image so that personal identification is employed using these ridges information. Fandong Zhang and Jufu Feng developed a model based on CNN and joint KNN-Triplet embedding. [18]

E. Generative Adversarial Network (GAN) based SR Model

Syeda Nyma Ferdous et al.[19] employed SRGAN for extracting features for detecting minutiae, ridge and pores to be enhanced so that personal identification is possible in all challenging situations but major hurdles of applying GAN model is computational complexity. Chi-En Huang et al. [20] employed residual GAN for enhancement. In their model, they employed residual network with attention module as generator and classification module will act as discriminator. Rafael Bouzaglo and Yosi Kellerc [22] developed generative

adversarial network with Resnet 50 as encoder, and convolutional decoder used for reconstruction. Masud An Nur Islam Fahim and Ho Yub Jungy [23] employed GAN based model for reconstructing good quality fingerprint image by applying skip connections on denoising auto-encoders and thereafter convolutional layers are appled for decoding. Amol S Joshi et al. [24] employed conditional GAN for deblurring input image followed by multiple discriminators. Gan based SR model requires training for both generator and discriminator separately and it takes longer time for reconstructing a better-quality image compared to other models, hence it is practically not feasible for employing to realtime applications. Mingzheng Hou et al.[25] employed GAN based network for generating good quality image. They applied SRResnet for upsampling the input image of appropriate size. Then, upsampled image was fed into an attention-based network for regenerating good quality image. Multiple discriminators were then applied to determine how well the regenerated image matched with ground-truth image. Since their model too complex, it is computationally expensive.

It is clear from the analysis of the previously mentioned SR models that there are a range of models for producing high-quality images from low-resolution data. All image processing applications require high-quality images in various scales to be supplied with less computations and minimal complexity in training and testing. Success rate of deep learning-based applications rely on learning strategy, availability of data-set, architecture of the model employed. Hence convolutional neural network architecture, a light weight neural network architecture with suitable number of layers, activation functions and optimizing functions is suitable to reconstruct images with higher PSNR and SSIM values with minimal training and testing computations.

III. Proposed Methodology

A. Model Architecture

In this model, initial step is pre-processing of input image (low-resolution image). In pre-processing stage, the gray scale finger print raw image is converted into unit 8 bit image. Then its pixel values are normalized to values between 0 and 1 by dividing pixel values by 255.

The preprocessed data is given to the deep learning model for training and testing. Architecture of proposed model shown in Fig. 1. In order to build a suitable model for enhancement of fingerprint images, we developed and tested 5 different behaviours of this architecture from scratch by changing number of filter (see Section V-A). Model details are explained in the Fig. 1. The input size of model is kept as 50×50 and the model predicted an output image of size 400×400. So, a resolution increasing factor of 8 times is achieved by this proposed system. The quality of generated HR image is verified and analysed using various standard techniques (see section V).

All five convolutional layers employed Relu as activation function and Adam as optimizing function with stride value as 1. After feature extraction and image reconstruction process, sub pixel convolution layer applied for up sampling LR image in to 8 times. Sub pixel convolution operates on 3 channels (3 sub pixels) of every pixel and combines 3 values for up sampling and regenerated image is very similar to ground truth image. Fig. 1 shows proposed SR model architecture with CNN and sub-pixel convolution operation.

B. Pixel Loss

During model training, the model weights are updated based on the custom pixel loss functions L_{pixel}. It is computed as per (1). In (1), GT represents ground truth image and SR image represents reconstructed image.

$$L_{pixel}(GT, SR) = \frac{1}{hcw}||GT - SR||_2^2 \tag{1}$$

Where h, c and w are the height, number of channels and width of the image.

IV. Experiment

A. Dataset

Proposed model trained from scratch using DIV2k data-set (https://data.vision.ee.ethz.ch/cv/DIV2K) and fine-tuned with publicly available standard fingerprint dataset - FVC2004 (http://bias.csr.unibo.it/fvc2004/databases.asp). This dataset is provided by the Biometric Systems Lab (University of Bologna), the Pattern Recognition and Image Processing Laboratory (Michigan State University) and the Biometric Test Center (San Jose State University). They provide three sets of finger print images with different resolutions and types. Table I shows the details of finger print dataset.

TABLE I. Dataset Details with Number of Images and Resolution

Data set	Number of images	Image size	Resolution
FVC 2004 DB1	240	640x480 (307K pixels)	500 dpi
FVC 2004 DB2	240	328x364 (11K pixels)	500 dpi
FVC 2004 DB3	240	300x480 (144K pixels)	500 dpi

B. Performance Evaluation

Well-known objective evaluation methods for measuring image quality include peak signal-to-noise ratio (PSNR), mean squared error (MSE), and structural similarity index (SSIM)[21]. This metric is defined as

$$MSE = \frac{1}{MN}\sum_{1}^{M}\sum_{1}^{N}(x_{i,j} - y_{i,j})^2 \tag{2}$$

Where $x_{(i,j)}$ represents original reference image and $y_{(i,j)}$ represents generated image and i and j are pixel positions of M N size image.

Peak signal-to-noise ratio (PSNR): PSNR is evaluated in decibels and is inversely proportional the Mean Squared Error. It is given as

$$PSNR = \frac{10log_{10}(2^n - 1)}{\sqrt{MSE}} \tag{3}$$

The higher values of PSNR denote the better quality of the reconstructed image.

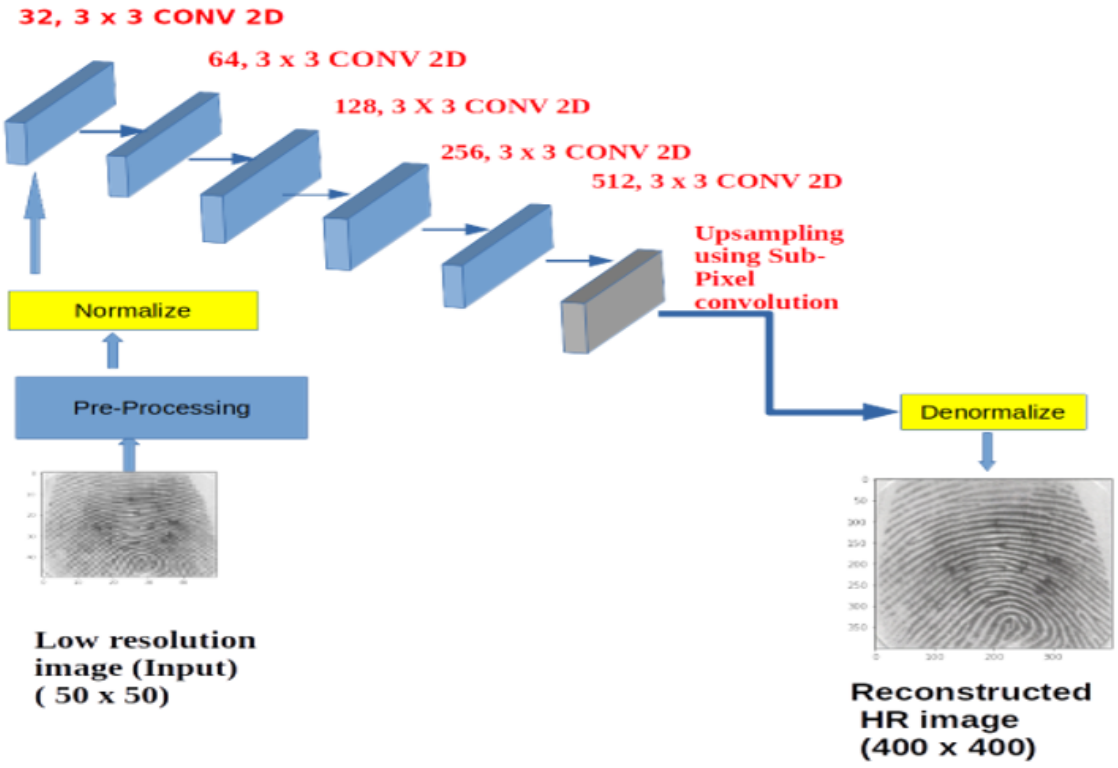


Fig. 1. Proposed Model Architecture.

TABLE II. MODEL BEHAVIOUR AND PERFORMANCE IN TERMS OF NUMBER OF TRAINABLE PARAMETERS, MODEL SIZE, PSNR, SSIM AND MSE VALUES, BY VARYING NUMBER OF FILTER CHANNELS & NUMBER OF CONVOLUTIONAL LAYERS

Case#	No of convolutional layers	Number of filter channels in each layer	Filter kernel size	Activation function applied	Optimizing function used	Mean PSNR, SSIM, MSE values	Number of trainable parameters and model size
1	5	Layer1 :8, Layer 2: 16 Layer 3: 32, Layer 4:64 Layer 5:128	3 x3	ReLu	Adam	PSNR: 8.4521 SSIM: 0.2475 MSE: 84.884	542,995 2.1 MB
2	5	Layer1 :16, Layer 2: 32 Layer 3: 64, Layer 4:128 Layer 5:256	3 x3	ReLu	Adam	PSNR: 19.45 SSIM: 0.687 MSE: 37.457	11,32,131 4.4 MB
3	5	Layer1: 32, Layer 2: 64 Layer 3: 128, Layer 4:256 Layer 5:512	3 x3	ReLu	Adam	PSNR :34.875 SSIM 0.9458 MSE: 22.04	2,897,923 11.13 MB
4	5	Layer1: 64, Layer: 128 Layer 3: 256, Layer 4:512 Layer 5:1024	3 x3	ReLu	Adam	PSNR: 35.788 SSIM : 0.9521 MSE: 18.967	87,795,87 33.5 MB
5	5	Layer1: 128, Layer 2: 256 Layer 3: 512 Layer 4:1024 Layer 5:2048	3 x3	ReLu	Adam	PSNR: 34.116 SSIM: 0.9501 MSE: 17.64	29,943,235 114.3 MB

Structural similarity index: SSIM measure similarity with greater accuracy and consistency than MSE and PSNR. It measures similarity between two images. It compares two images in terms of luminous, contrast and structure. The SSIM measure between two images x and y of size $N \times N$ is given as

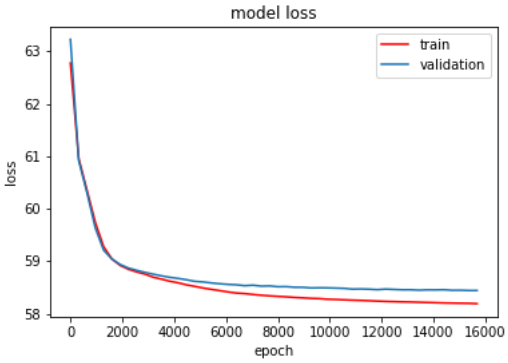


Fig. 2. Training and Validation Loss Curves.

TABLE III. PSNR VALUES OBTAINED DURING MODEL TESTING USING TEST SAMPLE IMAGES AND ITS COMPARISON WITH STATE OF ART METHODS EMPLOYED ON SAME DATA-SET

Image #	EDSR	WDSR	GAN Based model	Proposed Model
1	30.1	31.43	30.1	34.87
2	30.22	31.53	30.45	34.98
3	30.6	31.56	30.75	34.65
4	30.22	31.7	29.98	34.95

TABLE IV. MSE VALUES OBTAINED DURING MODEL TESTING USING TEST SAMPLE IMAGES AND ITS COMPARISON WITH STATE OF ART METHODS EMPLOYED ON SAME DATA-SET

Image #	EDSR	WDSR	GAN Based model	Proposed Model
1	25.36	27.73	29.87	22.05
2	25.55	27.43	29.1	22.1
3	25.22	27.1	29.43	21.98
4	25.62	26.98	29.1	22.01

$$SSIM(x_{(i,j)}, y_{(i,j)}) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (4)$$

Where C1, C2 are constants, μ_x and μ_y represents average of x and average of y. σ_x, σ_y represents standard deviation between ground-truth and regenerated images. SSIM values ranges between (-1,1).

In order to analyse the performance of the proposed model quantitatively we measured SSIM, PSNR and MSE values using reconstructed image with reference to the ground-truth image.

C. Training Strategies

The proposed model is trained and evaluated on a Nvidia Quadro T1000 4GB GPU based PC with 64-bits Windows Intel Xeon CPU at 2.60 GHz. Programs were written in Python language with Keras technology having TensorFlow as backend.

Input size is fixed as 50×50. Total there are 720 finger print images. We split the full dataset into train set and test data in the ratio of 9:1. So, 648 train data and 72 test data. Model is trained for 16000 epochs. During each epochs better model with lesser validation loss than previous is saved to hard disk using callback function of keras technology. So, after perfectly trained, best model with least validation loss is saved to hard disk for further testing and evaluations.

V. RESULTS AND DISCUSSIONS

Model trained for 16000 epochs, and got converged at 15680th epoch with validation loss of 58.4445. Fig. 2 shows the training and validation curves of model training stage.

Fig. 3 shows generated high resolution image of size 400×400 from low resolution input image of size 50×50. The generated HR image is visually compared with original ground truth image of size 400×400. In Fig. 3, generated images are visualized in a zoomed-in form just to visible the finger print patterns clearly. From the result figure, visually there is no any difference between the original ground truth image and generated HR image.

A. Research on Model Architecture

In order to build a suitable model for enhancement of fingerprint images, we developed and tested five different behaviours of this architecture from scratch by changing number of filter channels as shown in Table II. In each case we measured performance metrics quantitatively in terms of PSNR (Peak signal to noise ratio), SSIM (Structural similarity index), MSE (Mean squared error), Validation loss, Number of trainable parameters and Model size.

In first case, employed five convolutional layers and number of filter channel applied are 8, 32, 64, 128 and 256 accordingly. In this case, we obtained resultant image of poor quality in terms of different performance metrics mentioned above. In second case we adopted filter channels in each of the five convolutional layers are 16, 32, 64, 128 and 256 and obtained a result better than first case. In third case we employed convolutional layers with filter channels as 32, 64, 128, 256 and 512. In this case we obtained good quality image with significantly better value for performance metric like PSNR, SSIM, MSE, pixel-loss, number of trainable parameters and model size. In fourth case, we applied number of filter channels in each convolution layer as 64, 128, 256, 512, 1024. In this case, results obtained is slightly better than previous three cases except model size, which is a higher value. In last case, we adopted number of convolutional channels are 5 and number of filter channels employed are 128, 256, 512,1024, 2048. This case also we got better value in all performance metrics except model size and number of trainable parameters, which is a higher value than all other cases. From this comparison of five cases, we fixed third case as our proposed

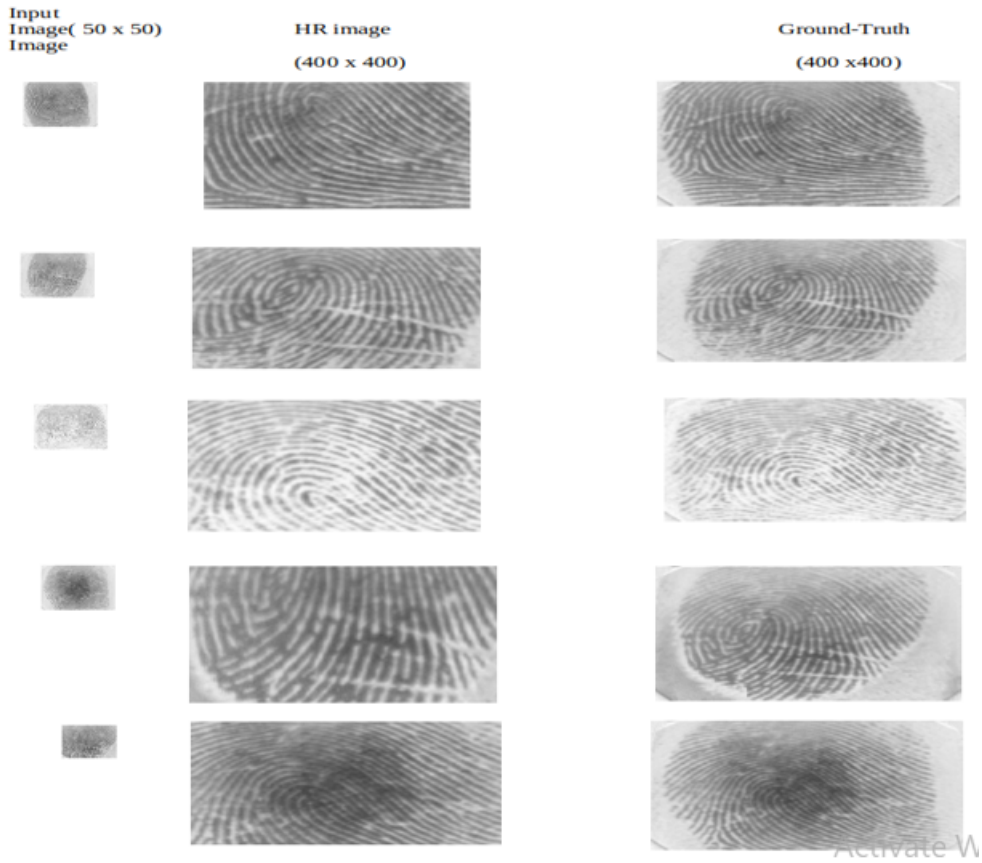


Fig. 3. Visualization of Some Input Images, Corresponding Generated Output Images and its Ground Truth Images.

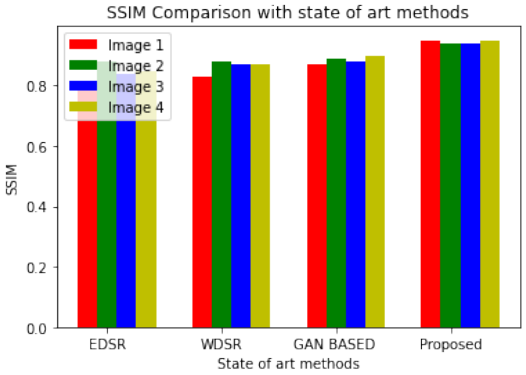


Fig. 4. SSIM Comparison with State-of-the-Art Methods.

model in which performance metrics like number of trainable parameters, model size and pixel loss are considerably less compared to other cases and PSNR, SSIM and MSE values are much better.

B. Comparison with State-of-the-Art Models

In order to analyse the performance of the proposed model quantitatively we measured SSIM, PSNR and MSE values using reconstructed image with reference to the ground-truth image.

Fig. 4 shows the comparison of SSIM index of proposed system with the state-of-the-art SR image generation techniques. For this comparison, we took four images and calculated SSIM values from the output values of other techniques like EDSR, WDSR, and GAN-Based technique. From this chart figure it is understood that proposed technique has got best performance than other technique in the perspective of SSIM index. Table III shows the comparison of proposed technique's PSNR with other techniques. Similarly Table IV shows the MSE values obtained during model testing using test sample images and its comparison with state of art methods employed on same data-set. From these two tables, it is understood that, proposed technique's performance is superior to other state-of-the-art methods.

VI. Conclusion

Fingerprints are used in many different applications like security control, law enforcement, smart phones, and criminal investigations. Over a lengthy period of time, the forensic community has used fingerprints as their most common biometric characteristic. Using a convolutional neural network, a light weight neural network for feature extraction and image reconstruction, followed by sub pixel convolution for upsampling, has been proposed as a novel architecture in this paper for resolution enhancement of low-resolution fingerprint images to high-resolution images. In this study, we looked at five different behaviours of this model, analysed model performance by altering the number of filter channels in each

of the five convolutional layers, and then fixed the model, which exhibits a notable improvement in performance metrics like PSNR, SSIM, MSE, model size, validation toss, and the number of trainable parameters when compared to other state of art methods. With no loss of the promising feature in the LR image, the proposed model enhanced the LR image eight times. Big data applications that analyse or compare fingerprint images for personal identification and verification confront significant challenges in meeting the necessary hardware and software requirements. In this case, the suggested model significantly contributes to improving the resolution of fingerprint images by taking a 50×50 input image and enhancing it eight times without losing any promising aspects, making it appropriate for real-time applications as well. Future work will build on this work by altering the model architecture through the use of a GAN-based network with perceptual loss while maintaining the computational viability of the model, making it appropriate for real-time applications as well.

REFERENCES

[1] Lidong Wang et al.,"Big Data Analytics in Biometrics and Healthcare". Journal of Computer Sciences and Applications, 2018, Vol. 6, No. 1, 48-55.

[2] T. Sivakumar et al.," An Approach to Reduce the Storage Requirement for Biometric data in Aadhar Project".ICTACT Journal on Image and Video Processing, February 2013, volume: 03, issue: 03.

[3] Sabyasachi Dash et al.,Big data in healthcare: management, analysis and future prospects". J Big Data, 2019, 6:54.

[4] Zaid Bin Mushtaq et al.,"Super Resolution for Noisy Images Using Convolutional Neural Networks". Mathematics 2022, 10, 777.

[5] Andrew Gilman et al.,"Interpolation models for image super-resolution".4th In ternational symposium on electronics design, test and applications. DOI 10.1109/DELTA.2008.104.

[6] C Ledig et al.,"Photo-Realistic single image super-resolution using generative adversarial network". 2017 IEEE conference on computer vision and pattern recognition.pp.105-114.

[7] Nasrollahi."Super-resolution a comprehensive survey". Machine vision and applications 25, 1423-1468, 2014.

[8] Ganchimeg. G and Leopold. H., "Fingerprint image enhancement using filtering Techniques." International journal of advance studies. 2019. 7(5):637-645

[9] Nouf Saeed Alotaibi."A new method to enhance fingerprint image reconstruction using deep boltzmann machine". International journal of intelligent engineering and systems, Vol.13, No.1, 2020.

[10] Andreea-Monica, et al., "A Fingerprint Matching Algorithm Using the Combination of Edge Features and Convolution Neural Networks", Inventions 2022, 7, 39.

[11] Z. Yang, Y. Xu and G. Lu, "Efficient Method for High-Resolution Fingerprint Image Enhancement Using Deep Residual Network", 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 2020, pp. 1725-1730.

[12] Seonjae Kim et al. "Single Image Super-Resolution Method Using CNN-Based Lightweight Neural Networks", Appl. Sci. 2021, 11, 1092.

[13] Yongliang Zhang et al., Slim-ResCNN: A Deep Residual Convolutional Neural Network for Fingerprint Liveness Detection", IEEE Access, PP(99) 1-1.

[14] Sandoval Veríssimo de Sousa Neto,"Finger print image enhancement using fully convolutional deep auto-encoders", Brazilian Journal of Development", vol 8, no 7, 2022.

[15] S. Saponara, A. Elhanashi and Q. Zheng, "Recreating Fingerprint Images by Convolutional Neural Network Autoencoder Architecture,"IEEE Access, vol. 9, pp. 147888-147899, 2021.

[16] Ajnas Muhammed and Alwyn Roshan Pais."A novel ingerprint image enhancement based on super resolution", 6th International Conference on Advanced Computing & Communication Systems (ICACCS). 2020.

[17] Ayushi Tamrakar, Neetesh Gupta." Low Resolution Fingerprint Image Verification using CNN Filter and LSTM Classifier", International Journal of Recent Technology and Engineering (IJRTE), Volume-8 Issue-5, January 2020

[18] Fandong Zhang and Jufu Feng, "High-resolution mobile fingerprint matching via deep joint KNN-triplet embedding", Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence .2017.

[19] Syeda Nyma Ferdous et al. "Super-resolution guided pore detection for fingerprint recognition ".25th International Conference on Pattern Recognition (ICPR) Milan, Italy, Jan 10-15, 2021.

[20] Chi-En Huang et al., "Super-Resolution Generative Adversarial Network Based on the Dual Dimension Attention Mechanism for Biometric Image Super-Resolution". Sensors 2021.

[21] C.Sasi varnan et al."Image quality assessment techniques in Spatial Domain", IJCST V ol . 2, I ssue 3, September 2011.

[22] Rafael Bouzaglo and Yosi Keller. "Synthesis and reconstruction of fingerprints using generative adversarial networks". 17 January 2022 Computer Science ArXiv

[23] Masud An-Nur Islam Fahim And Ho Yub Jung, "A Lightweight GAN Network for Large Scale Fingerprint Generation", January 2020 . IEEE Access. PP(99):1-1

[24] Amol S Joshi,"FDeblur-GAN: Fingerprint Deblurring using Generative Adversarial Network". 2021 IEEE International Joint Conference on Biometrics (IJCB)Aug 2021 Pages 1–8.

[25] Mingzheng Hou et al," Semi-supervised image super-resolution with attention CycleGAN", IET Image Processing. 2022;16:1181–1193.

Mobile Application: A Proposal for the Inventory Management of Pharmaceutical Industry Companies

Alfredo Leonidas Vasquez Ubaldo[1], Juan Andres Berrios Albines[2], Jose Luis Herrera Salazar[3],
Laberiano Andrade-Arenas[4], Michael Cabanillas-Carbonell[5]
Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú[1,2,3,4]
Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú[5]

***Abstract*—In recent years, the development of mobile applications has been evolving and becoming more and more frequent. This event is positive, since it plays an important role in facing and mitigating the multiple adversities that appear in the different existing sectors, such as business. On the other hand, it was detected that a little known problem that many companies in the pharmaceutical industry experience is poor inventory management, which causes countless consequences, generally of a negative nature. For this reason, in this work it was decided to make a mobile application prototype to face this problem. In this regard, the RUP methodology was used, along with various computer tools, in order to elaborate the prototype. Besides, as a data collection technique, surveys were made, which were subjected to expert judgment, in order to qualify the prototype. Likewise, very satisfactory results were obtained, concluding that the mobile application prototype that was developed complies with all the necessary conditions to mitigate the inventory management problems of pharmaceutical industry companies.**

***Keywords*—Inventory management; mobile application; pharmaceutical industry; prototype; RUP methodology**

I. Introduction

In the context of the pandemic that we had to live through, no sector was spared from Covid-19. The health sector was also hit hard by said pandemic [1],[2].Currently, having good inventory control is very important for any company, regardless of its types and items. This action seeks to ensure that the elements that are needed, such as raw materials [3], supplies and spare parts, are available in a timely manner, in optimal conditions and in their respective locations.

It is necessary to keep in mind that in a review of studies it was mentioned that the organization's inventory management involves decisions that include financing, promotion, supply and acquisition management. All of them have high risks and have a direct impact on the financial framework [4]. Besides, proper inventory management ensures availability and minimizes investment when materials are needed.

In the case of the pharmaceutical industry, inventory management is also essential, both from a financial and operational perspective. Efficient inventory management reduces procurement and transportation costs. Likewise [5], it maintains an effective stock of products to meet the demands of customers and prescribers.

It cannot be denied that efficient inventory management is crucial for the future of any business, as it becomes a key factor for profitability thanks to its multiple benefits. Among them we have, for example, that it will allow the company to have a timely control of all its products [6], as well as knowing at the end of the period a reliable state of the economic situation.

On the other hand, efficient inventory management will also avoid causing problems that threaten the safety and health of patients. This is because the system will allow the proper management of products, thus preventing patients from acquiring expired, falsified [7], deficient and/or damaged medicines.

In this context, inventory management systems are very convenient. Implementing, consolidating and effectively applying an inventory control system helps in the progress of any business, as well as improves the efficiency of its activities [8]. If companies and organizations use an inventory system, they will obtain many benefits, since the correct inventory management will improve their decision making.

Based on everything mentioned above, it is justified that this work is very important, since it aims to contribute to all companies in the pharmaceutical industry that choose to implement the proposed mobile application. In the same way, the objective of the research is to design a mobile application prototype, using the Figma tool. This in order to improve inventory management in companies belonging to the pharmaceutical industry so that they achieve many benefits, such as the reduction of losses due to expiration dates and the improvement of customer service.

To all this, it is also worth mentioning the content of the next chapters that make up this work. In Section II, works related to the topic are shown together with an analysis; in Section III, the methodology used is indicated together with the tools that helped in the elaboration of the application prototype; in Section IV, the development of the work is detailed; in Section V, the results obtained are presented; in Section VI, the results obtained are explained and compared with prior knowledge on the topic; and finally, in Section VII, conclusions are drawn and ideas for future work are provided.

II. Literature Review

As it was well emphasized before, in the present work it was decided to address the issue of designing a mobile

application prototype to improve inventory management in companies belonging to the pharmaceutical industry, since it was evidenced that it is a common problem today. For this reason, it was decided to search for scientific productions that are useful and related to the topic, in order to collect information on their observations, results, conclusions and other relevant aspects and learn from them.

In the first instance, in a work carried out by the authors [9], it was proposed to design and implement a pharmacy management system with a stock alert system, in order to improve accuracy and improve safety and efficiency in the pharmaceutical store. About it, in its conclusions it is indicated that the developed software allows new prescriptions and refills to be processed more quickly and easily; at the same time, this makes the work of pharmacists automated, allowing them to have more time to advise clients and thus prevent them from making medication errors.

On the other hand, in a work carried out by the author [10], it was proposed to design a mobile application for inventory management in a minimart. In the process, Waterfall was the methodology used, with the stages of analysis, design, coding and implementation. Besides, in its conclusions it is indicated that among the many benefits of this application is that it allows the management of articles to be easier.

From another approach, in a work carried out by the authors [11], it was proposed to develop a system to accurately manage consumable goods in storage. In the process, Systems Development Life Cycle (SDLC) was the methodology used. Likewise, the mobile application was developed using the Android system. Besides, in its conclusions it is indicated that this system has proven capable of reducing inventory access time by 80% and accurately tracking inventory compared to manual stock counting.

Meanwhile, in a work carried out by the authors [12], it was proposed to develop an information system for the hotel industry, in order to facilitate the control of data and inventory, orders and acquisitions and guarantee the follow-up of the cleaning process and consumption of materials as a whole. In the process, Design Science Research (DSR) was the methodology used. Besides, in its conclusions it is indicated this system has multiple benefits, among which it is mentioned that it allows data control and analysis to be carried out very easily.

From another perspective, in a work carried out by the authors [13], it was proposed to develop a mobile application for inventory management with sales prediction. In the process, regression analysis, typical of data mining, was used. Likewise, the mobile application was developed using the Android system. Besides, in its conclusions it is indicated that this application helps companies achieve greater social empowerment and development.

Similar to the previous one, in a work carried out by the authors [14], it was proposed to develop an inventory management system using the Rule of Association, in order to ensure that stores properly maintain their records and update your items in stock. In the process, the association rule, typical of data mining, was used. Likewise, AngularJS was used for the implementation of the system; PHP (Hypertext Preprocessor) for the backend of system development and database management; HTML (HyperText Markup Language) and CSS (Cascading Style Sheets) for system interface design; and NoSQL as the database engine. Besides, in its conclusions it is indicated that this system was very useful, since it allowed creating transactions, updating items in stock, keeping records, generating reports for decision-making and making stores more effective.

Taking into account the previous works investigated, it can be seen that there is a limitation in the use of software development methodologies, since they only appear in some articles. In addition, it is also evident that there is a limitation in the exploration of different platforms that exist to develop mobile applications.

III. Methodology

Poor performance and unreliability of applications are common factors that drastically affect their acceptance. In this regard, measures must be taken in terms of quality. For this reason, in the development of this work, it was decided to use a software development methodology, together with various computer tools, in order to develop the prototype of the inventory management mobile application for pharmaceutical industry companies.

A. The RUP Methodology

RUP (Rational Unified Process) is a software development methodology that is object-oriented. It is responsible for establishing the bases, templates and examples for each of the aspects and phases of software development. Furthermore, it combines aspects of the development process (such as defined phases [15], techniques and practices) with other development components (such as documents, models and manuals) within a unified framework.

This methodology is one of the most widespread and well-known among software development companies. It is based on the Unified Modeling Language (UML) and is characterized by being iterative and incremental, focused on architecture and guided by use cases [16]. Likewise, the goal of this methodology is to develop high-quality software, capable of meeting the needs of customers, within the costs and schedules planned for the project [17].

Having mentioned all of the above, the scheme of the RUP methodology can be seen in Fig. 1.

B. Phases of the RUP Methodology

The RUP methodology consists of four development phases, within which several iterations are carried out in order to satisfy defined criteria before embarking on another phase. In other words, if we want to advance to the second phase of the RUP methodology, we first have to meet all the criteria established in its first phase.

1) Inception Phase: This first phase is very short and focuses on achieving the feasibility of the project. To do this, it is necessary to establish the scope, identify current and future risks, propose an overview of the software architecture and develop the plan for phases and subsequent iterations with customers or stakeholders.

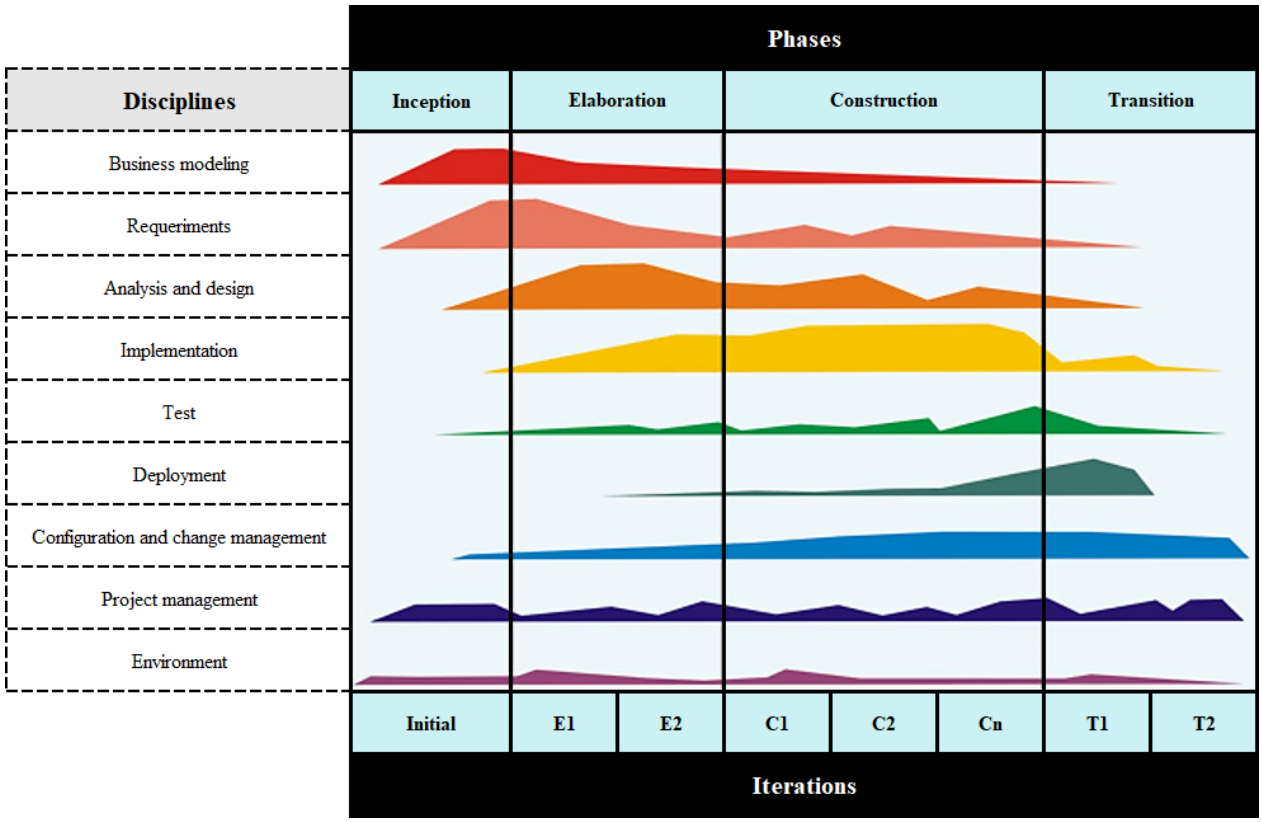


Fig. 1. Scheme of the RUP Methodology

2) Elaboration Phase: This second phase seeks to have a well-established base before moving on to the next. For this, it is necessary to select and elaborate the use cases [18], define the base architecture of the software, specify the selected use cases, develop the first analysis of the problem domain and design the preliminary solution.

3) Construction Phase: This third phase focuses on achieving the functionality of the software. For this, it is necessary to verify the pending requirements, manage the changes in relation to the evaluations made by the users and carry out the improvements.

4) Transition Phase: This fourth phase is the one that closes the project and seeks to ensure the availability of the software for end users. For this, it is necessary to carry out the final tests, adjust the errors and defects found, train users on the use of the software and provide the necessary technical support [18].

C. Elements of the RUP Methodology

A case apart from the four phases presented above, the RUP methodology is governed by four elements that work together and help to obtain the final result of the project.

1) Roles: It refers to the functions performed by each of the individuals or entities involved in the project. In this regard, it is worth mentioning that an involved party can play several roles, as well as the same role can be represented by several parties. In this sense, some roles [19], for example, could be that of a technical documenter, a software architect and a quality assurance.

2) Activities: It refers to the tasks that must be carried out by each of the individuals or entities involved in the project. In this regard, it is worth mentioning that each activity is assigned to a specific role. In this sense, some activities, for example, could be the elaboration of the use case diagram, the capture of software requirements and the performance of tests.

3) Artifacts: It refers to the products (in intermediate or final state) that originate during the various activities of the project and that are used to obtain the final result. In this regard, it is worth mentioning that the products capture information about the work carried out and transmit it. In this sense, some artifacts, for example, could be a document (such as the software architecture document), a model (such as the use case model) and an element belonging to a model (such as a class) [19].

4) Workflows: It refers to the sequence of activities that produce observable results of the project. In this regard, it is worth mentioning that all the roles, activities and artifacts that have been previously defined are integrated into the workflows. In this sense, some workflows [19], for example, could be a sequence diagram, a collaboration diagram and an activity diagram.

Having mentioned all of the above, the elements of the RUP methodology can be visualized in Fig. 2.

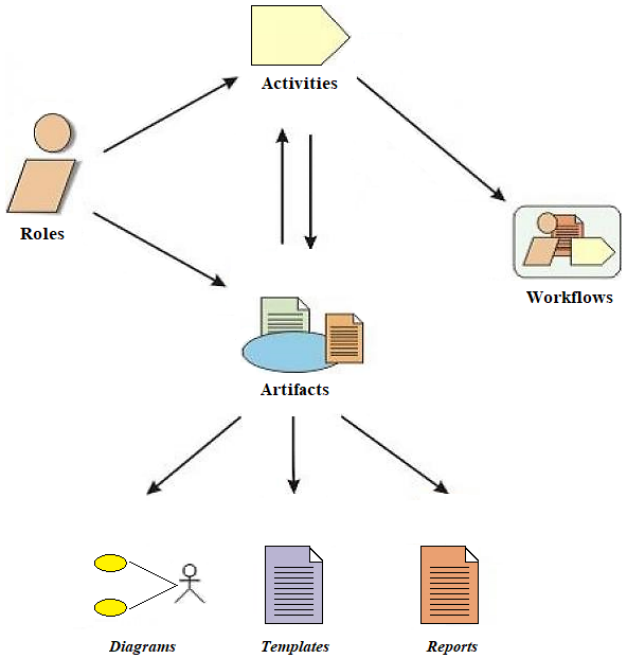


Fig. 2. Elements of the RUP Methodology

D. Technological Tools

1) Figma: It is used to create prototypes, through a development environment, which has an intuitive and easy-to-use interface. Figma allows designing user interfaces with excellent features in terms of design, prototype, collaboration, etc. In addition [20], it allows you to work collaboratively during the creation process and export the result in various formats such as PDF, PNG and JPG.

2) StarUML: It is an open source software modeling application based on UML standards [21]. It is flexible and easy to use.

3) Google Forms: It is a web-based application used to create forms for data collection purposes. Among its many benefits [22], its ease of handling when adding questions and answers stands out, as well as its ability to export the results in spreadsheets and statistical graphs.

IV. Study Case

During the development of the project, the activities of the phases of the RUP methodology that are most related to the development of the prototype of the mobile application for inventory management in a pharmaceutical industry company will be followed.

Based on the above, only the most necessary points of each of the phases of the RUP methodology were selected in order to achieve the final result that is the elaboration of the proposed prototype.

A. Phase 1: Inception

1) Scope of the Project: In this part, the scope of the project is established, which is related to its purpose. In other words, considering the purpose of this project, the need arose to raise some specific points, thus shaping the scope of the project. About it, the points that constitute the scope of the project are the following:

- Allow users to login.
- Register, modify and delete users.
- Register, modify, delete and query users.
- Register, modify, delete and consult providers.
- Register, modify, delete and consult products.
- Record sales.
- Generate sales reports.

2) Risks Associated with the Project: In this part, the risks associated with the project are identified to take them into account during the development of each of the activities because, as in any project, there are always situations that could occur and have a positive or negative impact on the final result. In this case, such uncertain events or conditions could be detrimental to scope and quality. About it, the risks associated with the project that were identified are the following:

- Inadequate choice of the technological tools to be used in the elaboration of the prototype.
- Ambiguous list of functional and non-functional requirements.
- Ambiguous definition of the roles of those involved in the project.
- Inaccurate rendering of Unified Modeling Language diagrams.
- Poor design of the prototype product of errors and details not considered.

3) Overview of Software Architecture: In this part, an overview of the mobile application architecture is proposed, which was done in order to provide a solid base to start modeling the system prototype. About it, the proposed architecture for this mobile application is presented in Fig. 3.

B. Phase 2: Elaboration

1) Software Requirements: In this part, the functional requirements and non-functional requirements of the system are defined. The functional requirements focus on the functionality of the system, since they are all those specific functions that the mobile application has. On the other hand, the non-functional requirements focus on the quality of the system, since they are all those specific attributes that the mobile application has. In this regard, these requirements are presented in Table I and Table II.

2) Business Use Cases: In this part, the business actors are established and the business use case diagram is designed. About it, these diagrams can be observed in Fig. 4 and Fig. 5.

3) Business Activities: In this part, the business activity diagram is elaborated. About it, this diagram can be observed in Fig. 6.

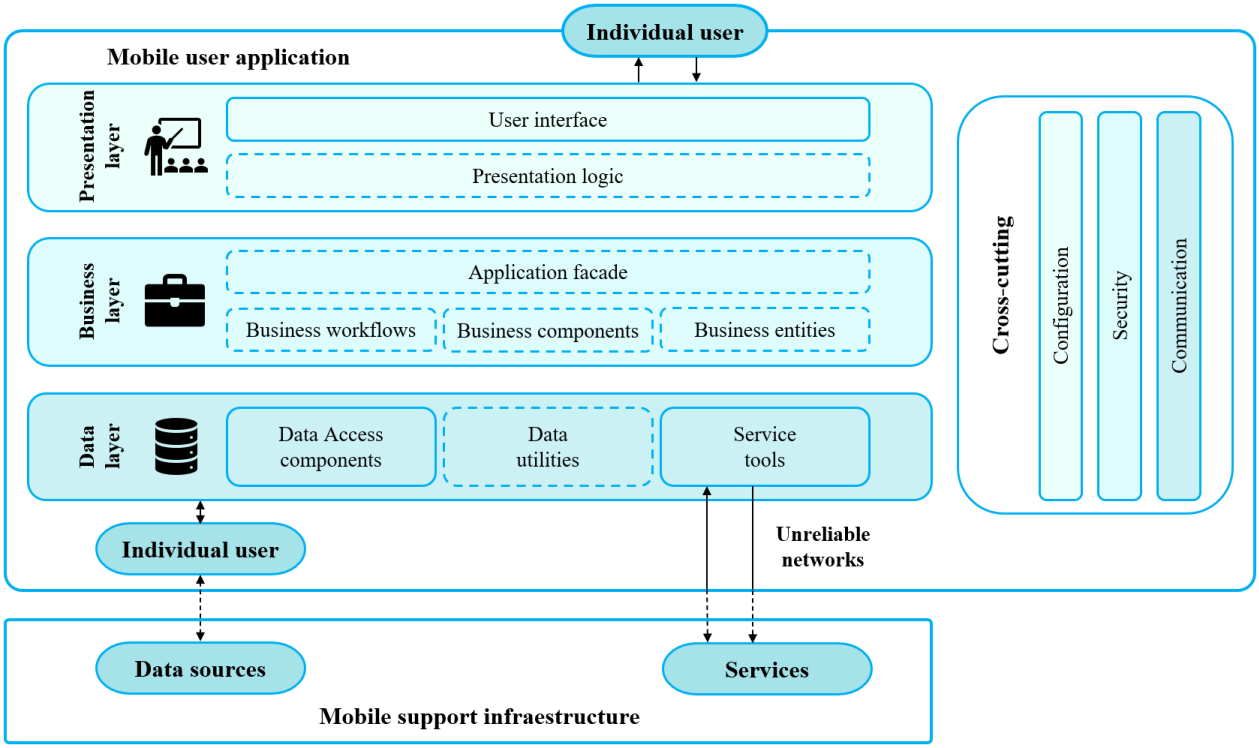


Fig. 3. Architecture of the Proposed Mobile Application

TABLE I. CAPTURE OF FUNCTIONAL REQUIREMENTS

Code	Description
FR01	The system will allow to validate the access of the users according to the entered data.
FR02	The system will allow to display a successful or failed login message
FR03	The system will allow to modify the login data of the users.
FR04	The system will allow to delete users.
FR05	The system will allow to register new users.
FR06	The system will allow to consult products.
FR07	The system will allow to modify of products data.
FR08	The system will allow to delete products.
FR09	The system will allow to register new products
FR10	The system will allow to consult providers.
FR11	The system will allow to modify of providers data.
FR12	The system will allow to delete providers.
FR13	The system will allow to register new providers
FR14	The system will allow to register the sale of products in real time.
FR15	The system will allow to generate sales reports.

TABLE II. CAPTURE OF NON-FUNCTIONAL REQUIREMENTS

Code	Description
NFR01	System learning time per user not exceeding 4 hours.
NFR02	Simple installation.
NFR03	Simple configuration.
NFR04	Disponibilidad de acceso 24/7 a todo el sistema.
NFR05	Safe and easy access.
NFR06	User authentication with a number of attempts not exceeding 3.
NFR07	Quick and easy navigation.
NFR08	Friendly and modern graphic interface.
NFR09	Ability to make changes and fixes.
NFR10	Ability to incorporate new functionalities.
NFR11	Compatibility with Android Studio and MySQL.
NFR12	Estándar de resolución 1440 x 2560: 560dpi.

4) System Use Cases: In this part, the system actors are established and the system use case diagram is designed. About it, these diagrams can be observed in Fig. 7 and Fig. 8.

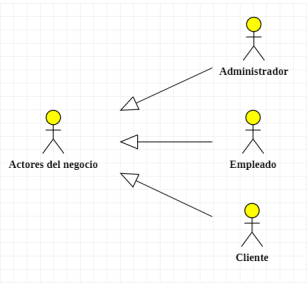


Fig. 4. Business Actors Diagram.

C. Phase 3: Construction

In this phase, the prototype of the mobile inventory management application is presented, which was developed with the Figma tool. This this application has two functionalities, these being the logistics area for employees and the sales area for customers. It is worth mentioning that the sales area was added as an added factor. Later, the most relevant interfaces of the application are shown.

In Fig. 9, it can be observed the charging interface, which appears when running the mobile application.

In Fig. 10, it can be observed the interface that appears after waiting for the application to load. This interface welcomes the

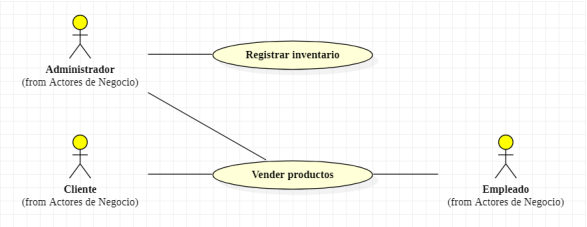


Fig. 5. Business Use Cases.

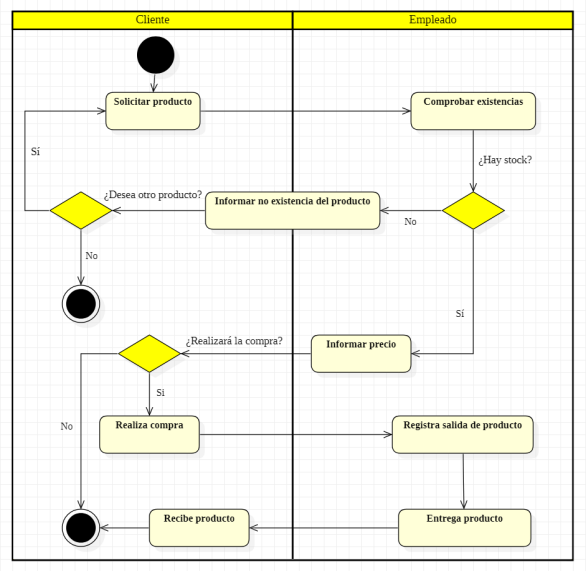


Fig. 6. Business Activity Diagram.

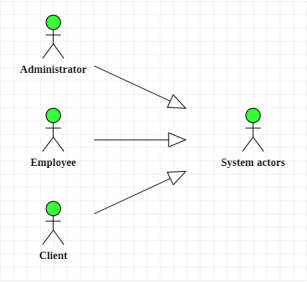


Fig. 7. System Actors Diagram.

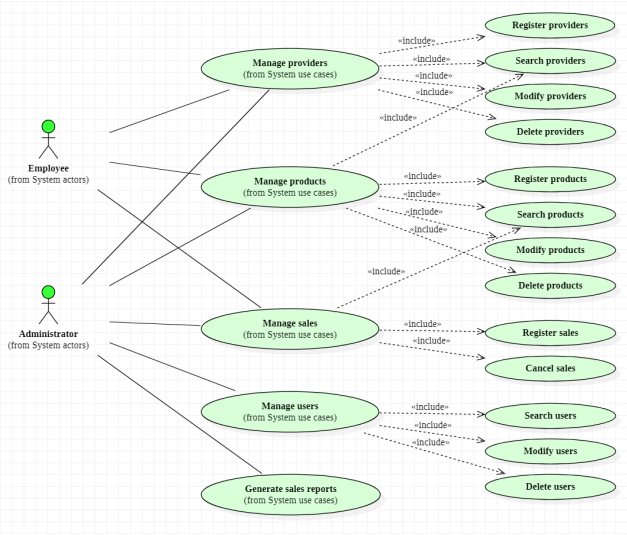


Fig. 8. System Use Cases.

user to the logistics area and asks him to enter his access data to enter. On the other hand, further down there is an option that directs the user to the entry interface to the sales area. The choice of which area to enter will depend on whether the user is an employee or a customer.

In Fig. 11, it can be observed the interface that appears after entering the access data. It shows the data of the employee who entered the system and gives the option to enter any of the windows you want.

In Fig. 12, it can be observed the interface that shows the providers part. It allows the user to contact them and register new ones.

In Fig. 13, two interfaces can be observed. The first interface allows you to register a new product. To do this, it will ask the user to enter their code and available stock. Also, it will ask the user to enter a comment about the product they are registering, as well as its expiration date. On the other hand, the second interface shows the list of products that are available.

In Fig. 14, it can be observed the interface that allows you to make sales reports, according to the date range that is entered. This interface shows the products sold, along with their respective units and prices. Also, it allows you to export the information.

In Fig. 15, it can be observed the interface that welcomes the user to the sales area. This interface shows the current products and offers. Also, it allows to search for products in the search engine.

In Fig. 16, it can be observed the interface showing the products part. In this interface the user can select any product he wants to buy.

In Fig. 17, it can be observed the payment interface, which allows you to make sales, asking if the payment method will be in cash or by card and if you want a bill or invoice. Also, this interface has the option to include an address for the delivery of the product.

D. Phase 4: Transition

It is important to validate that the elaborated prototype meets the necessary requirements of the end users and is free of errors and defects; otherwise, solutions will have to be found for the identified observations. For this reason, a validation of the prototype was carried out through expert judgment. Likewise, it is worth mentioning that the survey was prepared in Google Forms and contained fourteen (14) questions. About it, all this information can be observed in Table III.

On the other hand, it was sought that each expert, under their own criteria, qualify the proposed model. To do this, the survey responses were designed to be answered according to the Likert scale. Likewise, scores and percentages were

Fig. 9. Load Interface.

Fig. 11. Logistics Area Interface.

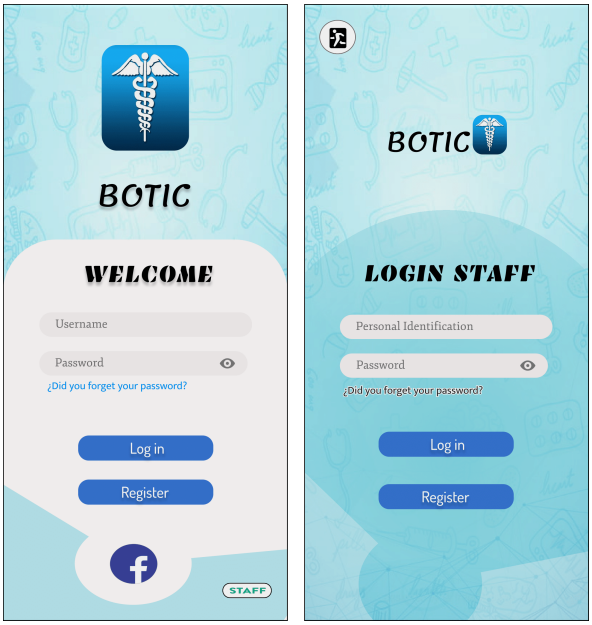
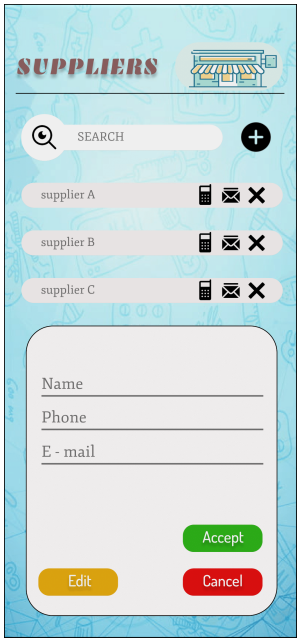


Fig. 10. Interfaces de inicio de sesión.



(a) N°5

Fig. 12. Provider Management Interface.

assigned. About it, all these data can be observed in Table IV.

V. Results

A. About the Case Study

In the inception phase, the scope of the project was established, focusing particularly on the operability of the system. Likewise, the risks associated with the project were identified, within which the prototype was tried to be free of errors and details not previously contemplated. Besides, an overview of the software architecture was proposed that served as a guide for the modeling of the system.

In the elaboration phase, the functional and non-functional requirements were captured. Likewise, five (5) UML diagrams were elaborated, these being the business actors, the business use cases, the business activities, the system actors and the system use cases.

In the construction phase, the prototype of the mobile

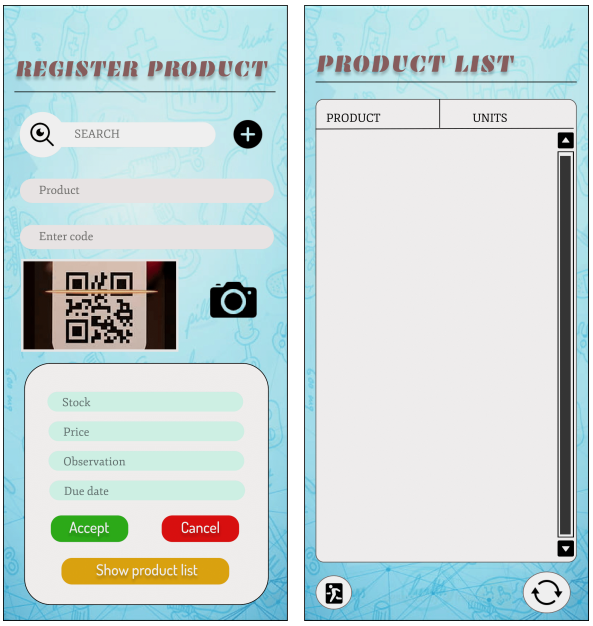


Fig. 13. Product Management Interfaces.

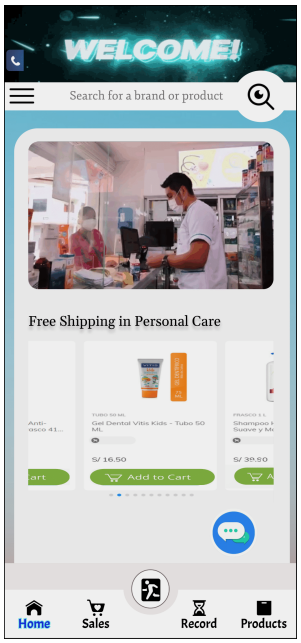


Fig. 15. Sales Area Interface.

Fig. 14. Sales Report Interface.

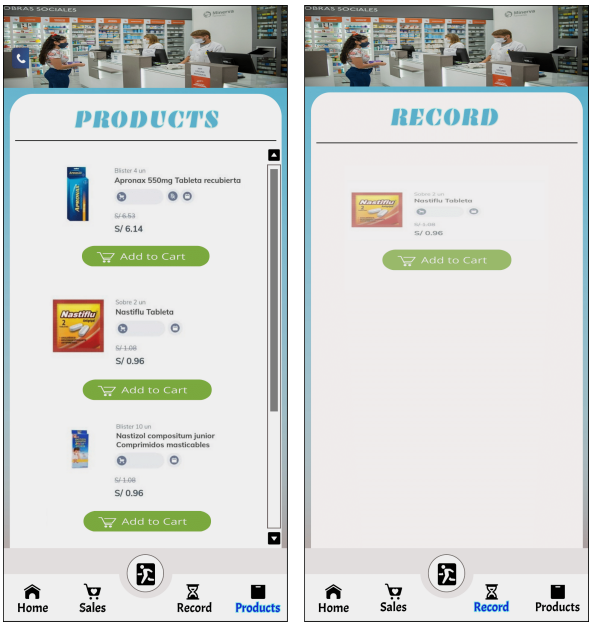


Fig. 16. Product Selection Interfaces.

inventory management application for the pharmaceutical industry was developed. About it, it is worth mentioning that the prototype consists of a wide variety of graphical interfaces, which allow users, products, suppliers and sales to be managed.

Finally, in the transition phase, surveys were carried out with the purpose of validating that the prototype is in optimal conditions and meets the needs of the user.

B. About the Survey

The survey model was elaborated with the Google Forms computer tool, which was very beneficial for the management of the data obtained.

Favorable results were obtained with respect to each of the survey criteria. About it, 96% of the total answers indicate that the respondents totally agree with the presentation of the application. Also, 92% of the total answers indicate that the respondents totally agree with the security, usability and functionality of the application. In short, all this shows that the application is optimal in terms of presentation, security, usability and functionality. For more details, all these data can be observed in Fig. 18.

Finally, it is necessary to mention that the survey

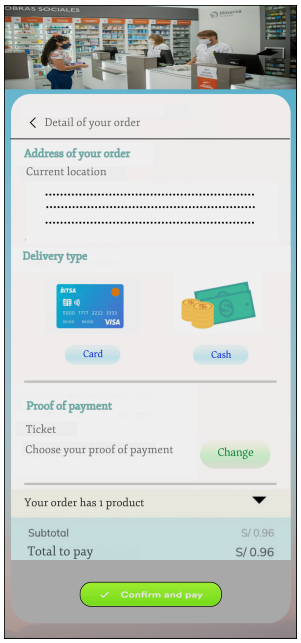


Fig. 17. Payment Method Interface.

TABLE III. SURVEY QUESTIONS

Criteria		Questions
Presentation	Q01	The interface of the application is modern.
	Q02	The interface of the application is friendly.
	Q03	The colors of the application stand out and contrast with each other.
	Q04	The graphics of the application pop and are timely.
	Q05	The windows of the application are well arranged.
	Q06	Indicate how much you agree with the presentation of the application.
Security	Q07	The access to the application is only for registered users.
	Q08	Indicate how much you agree with the security of the application.
Usability	Q09	The access to the application is easy.
	Q10	The navigation in the application is fast.
	Q11	Indicate how much you agree with the usability of the application.
Functionality	Q12	The application meets the needs of the user.
	Q13	The application will help improve inventory management.
	Q14	Indicate how much you agree with the functionality of the application.

application returned 93% in the final percentage. About it, all these data can be observed in Table V.

C. About the Methodology

As is well known, RUP is a methodology that focuses on planning and organizing a set of activities to turn user needs into software. The practices of this methodology are very common in large software projects; however, it is unknown what is the position of RUP compared to other software development methodologies. For this reason, a table was created to compare RUP with other software development

TABLE IV. ASSIGNMENT OF SCORES AND PERCENTAGES

Values	Scores	Percentages
Strongly disagree	1	20
Disagree	2	40
Neither agree nor disagree	3	60
Agree	4	80
Strongly agree	5	100

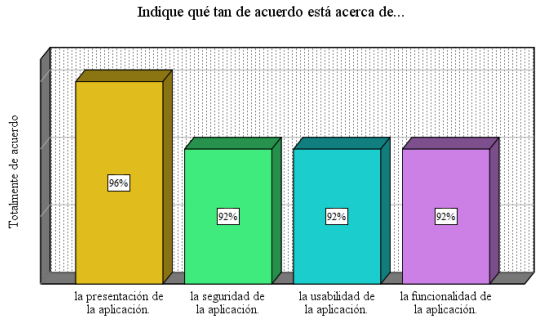


Fig. 18. Chart of Tables about the Opinion of the Respondents.

methodologies, in order to answer this question. About it, all this information can be observed in Table VI.

VI. DISCUSSIONS

During the literature review, works related to the development of mobile applications for inventory management were collected and analyzed. About it, it is worth mentioning that the works carried out by the authors [13] and [14] were very interesting. In both, data mining techniques were used in the methodological part and great results were obtained, demonstrating that there are other ways to develop software.

On the other hand, the literature review also allowed to compare the RUP methodology against other traditional software development methodologies, these being Waterfall and Incremental. Likewise, this comparison also involved some agile methodologies, these being Scrum, Kanban and Extreme Programming (XP). In this regard, it was identified that RUP is a very good methodology.

Now, regarding the surveys, it is worth mentioning that it is important that their application has yielded a high percentage of acceptance. This means that the prototype of the system is in optimal conditions and can be developed, since there is certainty that users will be highly satisfied with the application.

VII. CONCLUSIONS AND FUTURE WORKS

The prototype of the mobile inventory management application for the pharmaceutical industry developed meets the conditions of the users, given that the survey submitted to expert judgment was accepted by 93%. It was validated that the prototype has excellent presentation, security, usability and functionality. Therefore, it is concluded that there is certainty that the application will obtain great satisfaction from users and, therefore, can enter the development stage.

On the other hand, with respect to the case study, it is concluded that it was a good decision to use the RUP

TABLE V. SURVEY RESULTS

Questions	Experts										Percentages
	E01	E02	E03	E04	E05	E06	E07	E08	E09	E10	
Q01	5	5	4	5	4	5	5	5	4	5	94%
Q02	5	5	4	5	4	5	4	5	4	5	92%
Q03	5	5	4	5	4	5	4	5	4	5	92%
Q04	5	5	5	4	4	5	5	5	5	5	96%
Q05	5	5	5	4	5	4	5	5	4	5	94%
Q06	5	5	5	4	5	4	5	5	5	5	96%
Q07	5	5	4	5	5	4	4	5	4	5	92%
Q08	5	5	4	5	5	4	4	5	5	4	92%
Q09	5	5	4	5	4	5	5	5	5	5	96%
Q10	5	5	5	4	5	4	5	4	4	5	92%
Q11	5	5	5	4	4	5	4	4	5	5	92%
Q12	5	5	5	4	5	4	4	4	5	5	92%
Q13	5	5	4	5	4	5	4	4	5	5	92%
Q14	5	5	5	4	5	4	5	4	4	5	92%
Percentages	100%	100%	90%	90%	90%	90%	90%	93%	90%	99%	93%

TABLE VI. COMPARISON OF METHODOLOGIES OF SOFTWARE DEVELOPMENT

Criteria	Traditional methodologies			Agile methodologies		
	RUP	Waterfall	Incremental	Scrum	Kanban	XP
Proposals	It employs a set of activities necessary to transform user requirements into a system.	It linearly orders the different stages that must follow when developing the software.	It applies linear sequences in a staggered fashion as time progresses on the calendar.	It regularly applies a set of good practices to work collaboratively, as a team.	It applies a signaling system in which production tasks are displayed on demand by means of a series of cards.	It employs a set of techniques that provide agility, control, efficiency and flexibility in the development and management of the project.
Advantages	It allows early mitigation of high risks. It can be adapted and extended to meet the needs of any organization.	It provides the necessary tools to have clarity in the objectives from the beginning of the project. The costs and workload can be estimated at the beginning of the project.	It allows customers the opportunity to change requirements as components are added. It reduces the initial development time.	It allows to easily identify the objectives of each stage and the possible setbacks that may appear along the way. It manages the project in simpler and more manageable blocks, thus reducing the margins of error.	It allows variations in activities, thus ensuring that the product has the desired characteristics. It does not produce more than necessary, thus reducing waste.	It allows to save a lot of time and money. It has a very low error rate.
Disadvantages	The costs of the necessary team of professionals may not be covered on small projects. It may be unsuitable for use in small projects due to its high complexity.	It is difficult to go back and make changes. It delays tests until after completion.	It requires a lot of planning, both administrative and technical. It requires clear goals to know the status of the project.	It requires an exhaustive definition of the tasks and their deadlines. It requires that those who use it have a high qualification or training [23].	It is difficult to deliver on time on large projects. It is not implemented well in very long productive cycles.	It is difficult to keep track of what has been done. The commissions are very high in case of failure.

methodology, because it allowed the prototypes to be carried out in a systematic, orderly and coherent manner.

Finally, for future software development work, it is advisable to use other methodologies and other techniques related to data mining and technological trends.

REFERENCES

[1] F. Andrade-Chaico and L. Andrade-Arenas, "Projections on insecurity, unemployment and poverty and their consequences in lima's district san juan de lurigancho in the next 10 years," in *2019 IEEE Sciences and Humanities International Research Conference (SHIRCON)*, 2019, doi:10.1109/SHIRCON48091.2019.9024877, pp. 1–4.

[2] A. D. Rio-Chillcce, L. Jara-Monge, and L. Andrade-Arenas, "Analysis of the use of videoconferencing in the learning process during the pandemic at a university in lima," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.01205102

[3] K. Salas-Navarro, H. Maiguel-Mejía, and J. Acevedo-Chedid, "Metodología de gestión de inventarios para determinar los niveles de integración y colaboración en una cadena de suministro," *Ingeniare*, vol. 25, 2017.

[4] M. M. J. Basha, N. V.S, S. Wani, and V. Gogi, "Study of inventory management in pharmaceuticals: A review of covid-19 situation," *International Journal of Innovative Science and Research Technology*, vol. 5, pp. 366–371, 08 2020.

[5] A. Ali, "Inventory management in pharmacy practice: A review of literature," *Journal of Pharmacy Practice*, vol. 2, pp. 151–156, 09 2011.

[6] A. Ortega-Marqués, S. P. Padilla-Domínguez, J. I. Torres-Durán, and A. Ruz-Gómez, "Nivel de importancia del control interno de los inventarios dentro del marco conceptual de una empresa," *Liderazgo Estratégico*, vol. 7, 2017.

[7] R. Hidayat and I. Saleh, "The importance of inventory management in pharmaceutical practice," *Open Access Indonesia Journal of Social Sciences*, vol. 3, pp. 1–9, 06 2020.

[8] M. A. G. Segovia, S. B. R. Salvatierra, and R. Y. C. Y. Acebo, "Efficient inventory control," *RECIAMUC*, vol. 5, 2021.

[9] A. Baker, "Designing a computerized pharmacy management system with inventory stock alert system," *International Journal of Emerging Trends & Technology in Computer Science*, vol. 5, pp. 68–71, 10 2016.

[10] F. Darnis, "Mobile application for inventory control in a minimart," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 8, p. 101, 06 2017.

[11] R. Abdullah, K. Xiang, and M. I. H. C. Abdullah, "E-inventory management system using android mobile application at faculty of engineering technology laboratory stores." 2018.

[12] C. Kağnıcıoğlu and M. SEVER, "A new information system for inventory management in hospitality industry," *Journal of Business Research - Turk*, vol. 11, pp. 64–71, 02 2019.

[13] T. Tandel, S. Wagal, N. Singh, R. Chaudhari, and V. S. Badgujar, "Case study on an android app for inventory management system with sales prediction for local shopkeepers in india," *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 931–934, 2020.

[14] T. Oladele, R. Ogundokun, A. Adegun, E. Adeniyi, and A. Ajanaku, "Development of an inventory management system using association rule," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, p. 1868, 03 2021.

[15] P. Kruchten, *The rational unified process: an introduction*. Addison-Wesley Professional, 2004.

[16] J. L. Ávila Jiménez, *UF2406 - El cliclo de vida del desarrollo de aplicaciones*. Editorial Elearning, S.L., 2016.

[17] H. Engholm, *Análise e Design Orientados a Objetos*. Novatec Editora, 2017.

[18] R. L. Granados La Paz, *Despliegue y puesta en funcionamiento de componentes software. IFCT0609*. IC Editorial, 2015.

[19] P. Kroll and P. Kruchten, *The rational unified process made easy: a practitioner's guide to the RUP*. Addison-Wesley Professional, 2003.

[20] F. Design, "Figma: the collaborative interface design tool.(2017)," *Retrieved September*, vol. 17, p. 2017, 2017.

[21] E. Sutanto, *Pemrograman Android Dengan Menggunakan Eclipse & StarUML*. Airlangga University Press, 2020.

[22] J. Da Silva Mota, "Utilização do google forms na pesquisa acadêmica," *Humanidades & Inovação*, vol. 6, no. 12, pp. 371–373, 2019.

[23] R. Arias-Marreros, K. Nalvarte-Dionisio, and L. Andrade-Arenas, "Design of a web system to optimize the logistics and costing processes of a chocolate manufacturing company," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0120897

A Deep Neural Network based Detection System for the Visual Diagnosis of the Blackberry

Alejandro Rubio
Universidad Distrital
Francisco José de Caldas
Bogotá D.C., Colombia

Carlos Avendaño
Universidad Distrital
Francisco José de Caldas
Bogotá D.C., Colombia

Fredy Martínez
Universidad Distrital
Francisco José de Caldas
Bogotá D.C., Colombia

***Abstract*—Thanks to its geographical and climatic advantages, Colombia has a historically strong fruit-growing tradition. To date, the basis of its food and economic development in a significant part of its territory is based on a wide range of fruits. One of the most important in the central and western regions of the country is the blackberry, which is rooted not only from the economic and food point of view but also culturally. For the departments of Casanare, Santander, and Cundinamarca, this fruit is one of the primary sources of income, rural employment, and food supply and income. However, small and medium farmers cultivate without access to technological production tools and with limited economic capacity. This process suffers from several problems that affect the whole plant, especially the fruit, which is strongly influenced by fungi, extreme ripening processes, or low temperatures. One of the main problems to be dealt with in its cultivation is the spread of pests, which are one of the causes of fruit rot. As a support strategy in producing this fruit, the development of an embedded system for visually diagnosing the fruit using a deep neural network is proposed. The article presents the training, tuning, and performance evaluation of this convolutional network to detect three possible fruit states, ripe, immature, and rotten, to facilitate the harvesting and marketing processes and reduce the impact on the healthy fruit and the quality of the final product. The model is built with a ResNet type network, which is trained with its dataset, which seeks to use images captured in their natural environment with as little manipulation as possible to reduce image analysis. This model achieves an accuracy of 70%, which indicates its high performance and validates its use in a stand-alone embedded system.**

***Keywords*—Automatic sorter; blackberry; deep neural network; fruit handling; image analysis**

I. Introduction

Colombia is a country that has been characterized by its great variety of fruits and fauna [1]. Its diversity and climatic stability, as well as its great cultural variety, have made it possible to define different foods that provide the necessary nutrients with an unparalleled flavor [2]. One of the most known and used fruits in daily life is the blackberry because it has an excellent flavor and an endless number of preparations [3], [4], [5].

Blackberry cultivation is mainly carried out by small family businesses without large investments in technology. These farmers strive to maintain the quality of the fruit [6], however, it is difficult to sustain it in planting with a size of approximately two hectares. Also, the process is further complicated by having to use chemicals to care for the plant, which always has an impact on the quality of the final product [7], [8].

An autonomous artificial sorting system can perform effective segmentation from images of blackberry in its different growth processes. Such a system, easy to use and very low cost, can support the production and commercialization of the fruit in its different stages [9], [10]. This type of system should consider a robust and low-cost portable design [11], [12], [13]. In this sense, the use of an identification algorithm capable of operating in real-time on low-cost embedded hardware is essential [14], [15], [16].

Annually Colombia produces 137,999 tons of blackberry. About 55% of the production stays in Colombia for fresh consumption (supermarkets and marketplaces) and less than 1% is exported [17]. Between 2015 and 2018, 12.8 tons of blackberry were exported, and it is known that the department of Cundinamarca is one of the largest producers, followed by Santander. These have an annual production of 26% and 17% respectively. Blackberry, despite its unique flavor, requires special processes to maintain its quality and avoid premature fermentation. Blackberry production costs are estimated at USD 2,410, which is divided into four main activities: land preparation, planting, harvesting, and inputs. However, for each ton of blackberry it is estimated that the value paid to the farmer is USD 400, which means that for every 14 million, a gross profit of USD 1,542 is obtained [18], [19].

Blackberry exports in Colombia are less than 1% [20]. However, this is one of the crops that generates the highest permanent income due to the domestic market [21]. Twenty percent of the product remains in the agro-industry for the production of juices, and jams, among others, while the remaining 55% remains for fresh consumption. This indicates that about 75% of the production remains for domestic trade, and is distributed in the largest distribution centers of agricultural products in the country. For example, Corabastos in the Colombian capital handles 36% of production, followed by the city of Bucaramanga with 20% in 2019. Its economic importance is reflected in the domestic market due to its great demand, and this, in turn, generates jobs and economic resources for the Colombian population. Therefore, it is estimated that 1,900 new hectares of blackberry crops are established annually. These crops can be affected by pests, which can end up destroying the entire crop if they are not properly cared for or do not have the necessary resources [22].

Since blackberry is a perishable food that grows in the field,

and its characteristic of food for pests such as flies, thrips, mites, and even the so-called fruit worms [23], it is essential to perform a quick and efficient characterization of the affected fruit to reduce possible damage and losses. These pests cause the fruit to be damaged, and this in turn is spreading throughout the crop, i.e., will produce losses to the farmer. For this reason, they implement various chemicals to eradicate them, which means that the blackberry has a large amount of them. An automatic grading system capable of being used by the farmers would increase their production capacity and product quality [24], [25].

The blackberry is a product with broad social roots that benefits the Colombian population from a nutritional and economic point of view. The blackberry industry provides employment in economically depressed areas, and at the same time has great nutritional value. Although it is not a product that brings income to the country as an export product, it is very important for the domestic economy in many regions of the country. It is a product that should be consumed fresh because it has a characteristic of rapid fermentation that considerably reduces its quality. Also, to maintain the quality of the crops, certain chemicals are used to eradicate pests, therefore it is necessary to invest in technology to facilitate and reduce production costs while ensuring the quality of the fruit [26], [27]. An automatic recognition system allows keeping control of the fruit throughout the production and commercialization process, which provides added value to the process and reduces costs [28].

In current fruit production processes, there is no equipment or tools similar to the one proposed in this research [29]. Previously, fruit inspection systems for harvesting have been implemented with shallow success, particularly for using traditional image processing strategies for sorting, which is very unreliable in actual field applications.

The paper is structured as follows. Section II describes the problem under study and the possible profile of the required solution. Section III presents the methodological development followed for the solution, giving details of each section of the system, from the input dataset and its manipulation to the desired output categories, including the architecture designed for the model. Section IV summarizes and discusses the results achieved by the model, using metrics and discussing their scope. Finally, Section V concludes our paper by highlighting the strategies used and the most important findings.

II. Problem Statement

The objective of this research is to develop an autonomous classification model based on images to determine the conditions of blackberry fruit throughout the harvesting and marketing process. It is intended to use this model for the construction of a low-cost embedded system that is easy to use with farmers. With this system, the farmers will be able to identify in time the problems of the fruit in their harvest, reducing negative effects on production. The handling of this fruit must comply with certain processes that guarantee its quality, i.e., avoid the proliferation of pests and diseases that attack the crop, as well as maintain its growing conditions. Changes in these processes often affect the plant from the root to the fruit, potentially lowering the volume of the harvest. When one plant gets infected, the damage quickly spreads to other plants, generating large losses for growers. These crop problems are not only due to the misuse of insecticides or soil, but the weather plays an important role. Low temperatures produce frosts that influence the weakening of the plant and therefore facilitate the incorporation of pests and diseases.

Pests and diseases are not the only problems that cause damage and losses in this crop, an extreme ripening of the blackberry causes fermentation, making it impossible to consume. Blackberry is consumed fresh, inadequate handling and storage processes affect the quality of the fruit, an effect that also spreads rapidly throughout the product.

As a solution strategy, we propose an image categorization model based on a convolutional network. We propose the use of a ResNet (Residual Neural Network) trained and tuned with a proprietary dataset to identify three fruit states: ripe fruit (category 0), unripe fruit (category 1), and rotten fruit (category 2). This architecture is chosen due to its high performance in similar problems, and its small size which facilitates its use in a low-cost embedded system. The aim is that such a system can identify possible problems in the fruit early, before spreading and infecting the whole crop (harvest). The general model of the proposed system is shown in Fig. 1.

III. Methods

In the preliminary performance tests, the ResNet model obtained the highest average and category performance. The ResNet version of 50 layers deep (ResNet-50) was implemented, seeking the smallest possible size for the model. The great advantage of this model is that it reduces the depth by sending information forward, skipping layers, and improving at the same time the learning capacity [30]. The coding of the model was performed in Python 3.8.5 with support for numpy 1.19.2, scikit-learn 0.23.2, scipy 1.5.2, OpenCV 4.4.2, and matplotlib 3.3.2.

One of the essential features of our categorization model is that it must be able to produce results with images captured in real environments, such as those that a farmer could capture with his cell phone. In this way, the system would be easy to manipulate in real environments, and the categorization model would produce reliable results. To test this argument, we built our dataset from images supplied mainly by local farmers and experts in fruit handling. In addition, this database was supplemented with public images of fruit conditions. The images were not manipulated to extract or remove information from them, but the leading actor was always the fruit.

The training of the convolutional model was performed according to the following characteristics:

- **Dataset**. We built our balanced dataset with 100 images in each category (300 images in total, Fig. 2). The images in each category correspond to states identified by fruit condition experts. According to the performance of the system, this dataset can be increased if this increases the categorization performance. The possibility of continuous updating of the database and online training is proposed.
- **Manipulation of the Dataset**. The image processing was done with OpenCV. The images were randomly

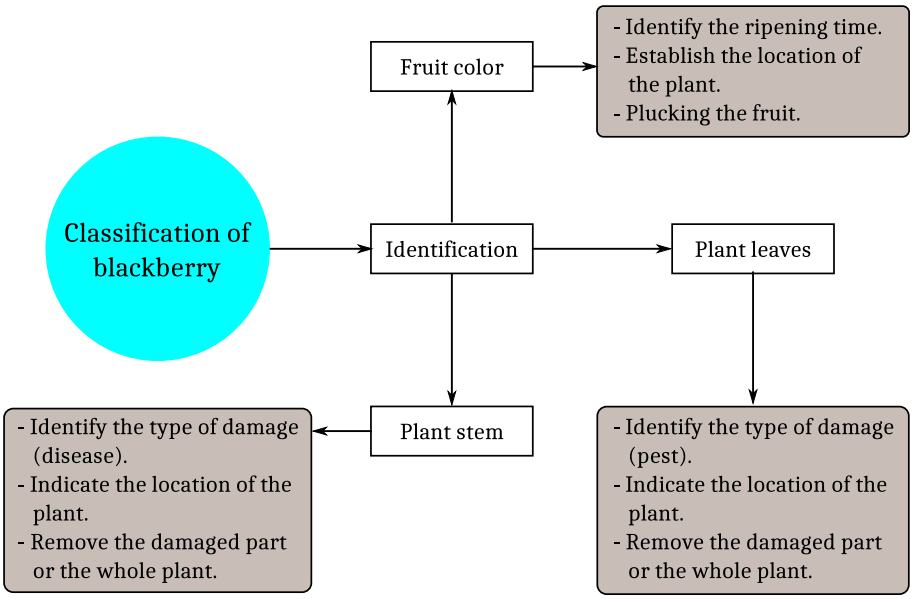


Fig. 1. General Diagram of the Classification and Control Scheme of the Blackberry Plant.

shuffled and scaled to a size of $32'times32$ pixels to reduce the complexity of the code. This size is conditional on the performance of the model, but in initial tests, it showed a high capacity to retain features while considerably reducing the processing power required on the embedded hardware. The images were processed in RGB matrices and normalized to the working values of the convolutional network.

- **Training and Validation**. To build the model, training is performed with 70% of the dataset (random selection of images), while the remaining 30% is used for validation. The optimization is performed with the Stochastic Gradient Descent (SGD) function, and the loss calculation is performed with the Categorical Crossentropy function. The training was performed over 30 epochs, and at each step, the accuracy and Mean Squared Deviation (MSD) error values produced for both training and validation data were controlled. The number of epochs was defined according to the learning capacity of the model, avoiding over-fitting.

- **Model Assessment**. The performance of the model was evaluated using three metrics: Precision, Recall, and F1-score. These metrics were calculated for the validation images in each category of the model, as well as its average behavior. The convolution matrix was also used to identify problems related to false positives and false negatives.

We use a ResNet-50 model looking for architecture with high performance, but at the same time suitable for propagation on embedded hardware (Fig. 3). This deep network is inspired by cells of the pyramid of the cerebral cortex to form equivalent functional blocks. This functionality is achieved using jumping connections to send information forward layers of the structure. It is thanks to these jumps that this deep model manages to avoid gradients that fade away during training, which is what gives it its high performance. This structure makes sense when all the intermediate layers are linear or are superimposed on the nonlinear layer, otherwise, the re-use of weights in the forward layers would not make sense. The ResNet-50 model has five stages, each with a convolution and identity block. Each convolution block contains three convolution layers, as does each identity block. With this structure, the model has a little more than 23 million parameters to be adjusted.

IV. Results and Discussion

The capacity and performance of the model were evaluated throughout the training with both training and validation data. Accuracy (Fig. 4) and the behavior of the loss function (Fig. 5) were calculated throughout each epoch. The Accuracy of a classification model indicates the number of correct predictions for the total number of input images. In Fig. 4 the red curve shows the Accuracy behavior for the training data, while the green curve shows the same metric for the validation data. The Accuracy of the training data remains always high, while the behavior of the validation data remains very poor, at least until epoch 13, from which the performance of the validation data increases according to the model fit.

Fig. 5 shows similar behavior. The red curve again shows the behavior for the training data, while the green curve shows the error produced by the validation images. Although in both cases there is error reduction, the initial model had a good performance for the training data, and the most marked reduction in error is observed in the validation data. Without wishing to cause training bias, the model was tuned to produce better behavior for unknown images. Again, the reduction in error becomes more marked from epoch 13 onwards.

The confusion matrix of the model allows to observe graphically the overall performance, as well as in each category, and allows for calculating the Precision, Recall, and F1-score metrics (Fig. 5). The matrix uses a heat grading that assigns

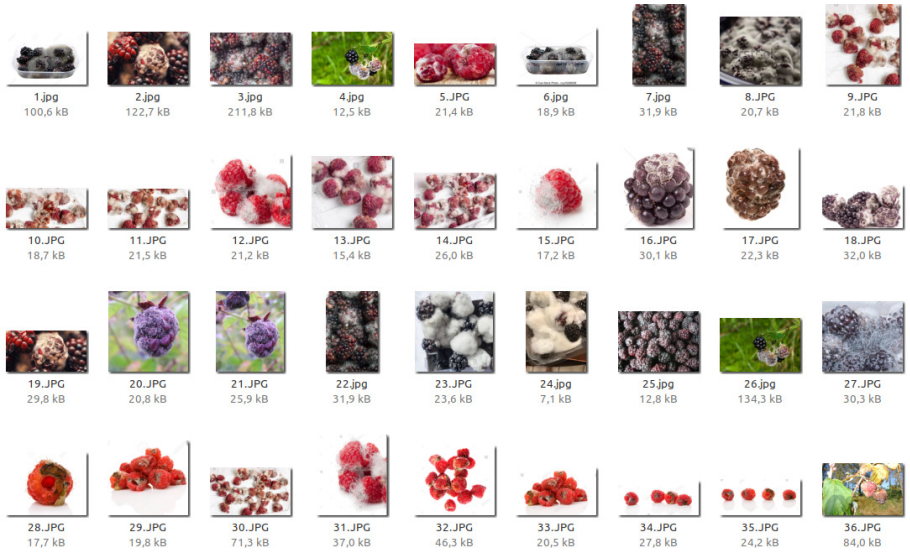


Fig. 2. Sample Dataset used in Category 2 (Rotten Fruit).

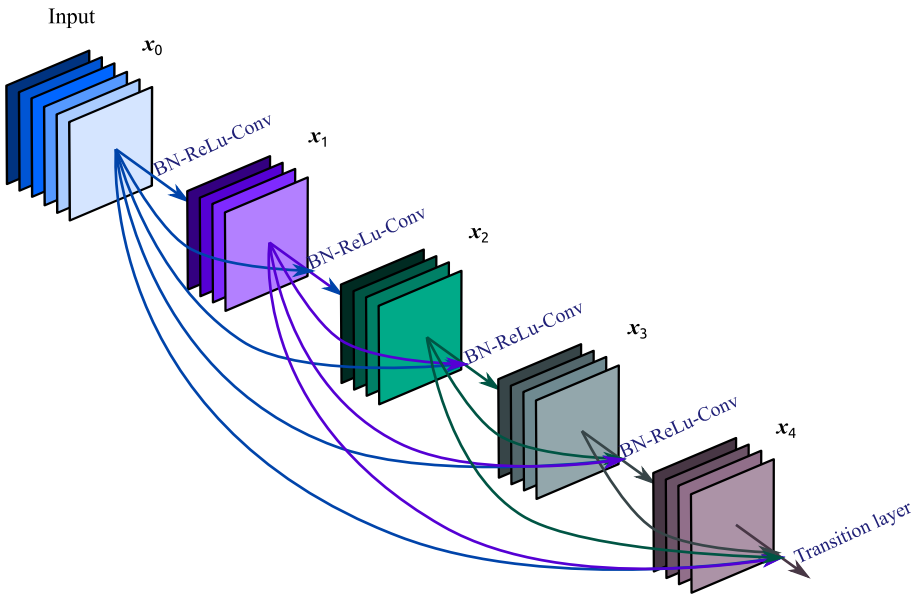


Fig. 3. ResNet-50 CNN Architecture.

light colors to high values, and dark colors to low values. The main diagonal in Fig. 6 with light colors, and the dark colors above and below it, show high categorization performance. Specifically, it is seen that in the ripe fruit category, 21 of the images are correctly classified, with very low percentages of false positives and false negatives. In the category of immature fruit, it is seen that it had more difficulties in the classification, since it only classified 18 of these, and in the category of rotten fruit it had the best performance since it was able to correctly classify 27 of the images.

We also calculated the Precision, Recall, and F1-score metrics for the model using the validation data (Table I). The accuracy of the model indicates how good the model is at placing the correct images in a given category, i.e., how many of those placed in a category belong to it. Our model obtained an accuracy of over 68% in all categories and an average value of 76%. Although not perfect, the values are more than good for the proposed development. On the other hand, the Recall shows how many of the positive ratings in each category belong to that category. In category 1, as indicated in the confusion matrix, we have a relatively low value (51%), but the values for the other categories exceed 84%, which speaks very well of the model. The average value for Recall is 75%. Finally, F1-score corresponds to a weighted average of the first two metrics, so it combines their qualities. The model obtained a value per category of over 75% and an overall average value of 73%.

This categorization model is being evaluated on the

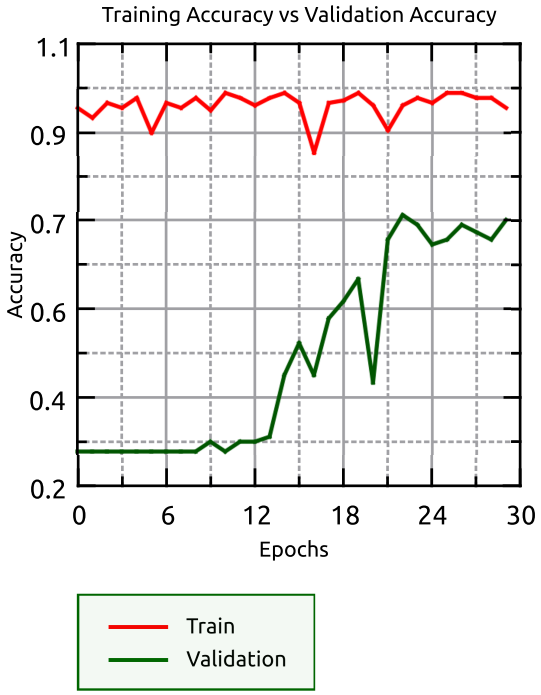


Fig. 4. Accuracy for Training and Validation Data throughout the Training Process.

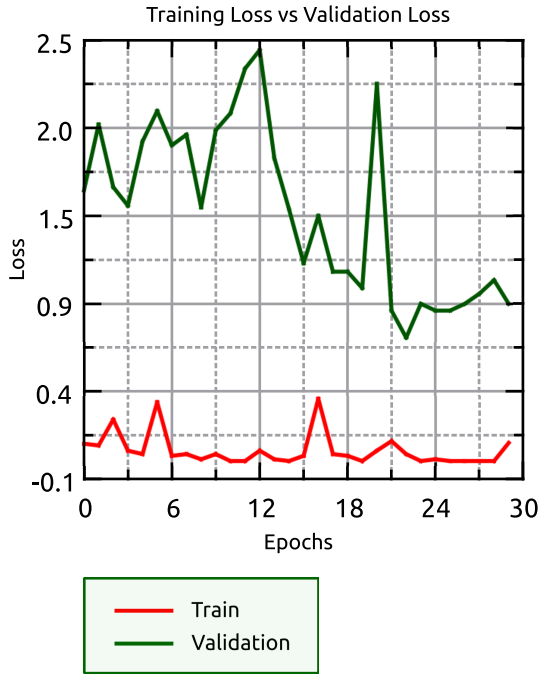


Fig. 5. Loss for Training and Validation Data throughout the Training Process.

TABLE I. SUMMARY OF THE BEHAVIOR OF THE MODEL METRICS

	Prescision	Recall	F1-score	Support
0	0.68	0.84	0.75	25
1	0.90	0.51	0.65	35
2	0.69	0.90	0.78	30
Accuracy			0.73	90
Macro avg	0.76	0.75	0.73	90
Weighted avg	0.77	0.73	0.72	90

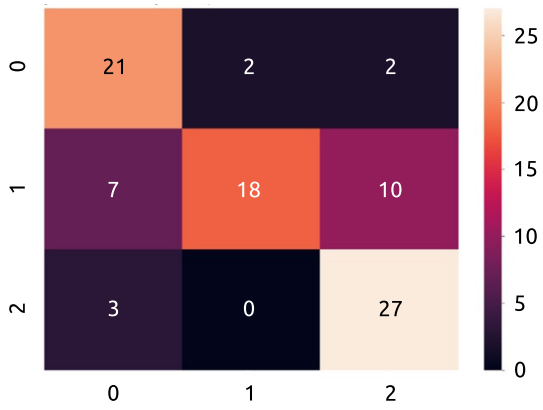


Fig. 6. Model Confusion Matrix.

DragonBoard™ 410C development board from Arrow Development Tools. This system is powered by a 64-bit ARM®Cortex™ A53 Quad-core processor and 1 GB of RAM. The system has been configured with Debian Linux, and in initial testing has demonstrated high real-time performance. Further evaluation will be conducted in the future development of this research.

Compared to traditional strategies for the identification and categorization of fruits, our proposal presents excellent advantages for implementation in real systems, not only because of the high performance reported by the model but also because of the possibility of working with images with minimal or no previous processing, which allows farmers to manipulate the tool and produce reliable results immediately directly. This advantage is further enhanced by the lack of the need for complex, high-cost hardware, which allows the development of prototypes on small embedded systems.

V. CONCLUSION

This paper presents an image categorization model of blackberry fruit as a strategy for the construction of an autonomous identification system to support the process of cultivation and marketing of the fruit. The ResNet-50 convolutional neural network was selected for the development of the model due to its high performance and small size. This network was trained, tuned, and validated with a proprietary dataset separated into three categories coinciding with the state of the fruit. The code was developed in Python with Keras and TensorFlow support, and a model with good performance suitable for embedded applications was generated. During training, the Categorical Crossentropy function was used as a loss function, and the Stochastic Gradient Descent function was used as an optimizer. The evaluation of the model was performed with the Precision, Recall, and F1-score metrics, as well as the confusion matrix of the model. According to the results with the validation data (Precision of 76%, Recall of 75%, and F1-score of 73%) the model has adequate

performance for the development of the prototype, also, the behavior throughout the training allows intuiting that it is still possible to increase the learning and performance of the model. It is expected that the system can be used to perform on-site fruit sorting, allowing the early identification of rotten fruit, reducing damage at harvest, and handling the product. Future directions of the project are oriented to allow real-time updating of the database and evaluation of the system on hardware prototype.

ACKNOWLEDGMENT

This work was supported by the Universidad Distrital Francisco José de Caldas, specifically by the Technological Faculty. The views expressed in this paper are not necessarily endorsed by Universidad Distrital. The authors thank all the students and researchers of the research group ARMOS for their support in the development of this work.

REFERENCES

[1] W. Sánchez and V. Guerrero, "Ecology, construction, and innovation: An alert towards change," *Tekhnê*, vol. 15, no. 2, pp. 45–58, 2018.

[2] W. A. Cardona and M. M. Bolaños-Benavides, "Manual de nutrición del cultivo de mora de castilla (rubus glaucus benth.) bajo un esquema de buenas prácticas en fertilización integrada," *Agrosavia*, 2019.

[3] J. Barrera, I. Moncayo, D. Cruz, J. Pinzón, J. Gómez, and H. Moreno, "Art: Caracterización fenotípica y organoléptica de mora (rubus spp) cultivadas en el área metropolitana de bucaramanga, santander," *RI-UTS*, 2020.

[4] P. A. Martínez, J. L. Ramirez, and V. Q. Castaño, "Formulación y evaluación fisicoquímica de jugo de mora (rubus glaucus benth) enriquecido con calcio y vitamina c," *Biotecnología en el Sector Agropecuario y Agroindustrial*, vol. 18, no. 1, pp. 56–63, 2020.

[5] B. L. Moreno and Y. A. D. Oyola, "Caracterización de parámetros fisicoquímicos en frutos de mora (rubus alpinus macfad)," *Acta Agronómica*, vol. 65, no. 2, pp. 130–136, 2016.

[6] E. Sánchez-Betancourt, M. C. García-Muñoz, J. Argüelles-Cárdenas, V. Franco-Flórez, and V. Núñez, "Atributos de calidad de frutos en diez genotipos de mora colombiana (rubus glaucus benth.)/fruit quality attributes of ten colombian blackberry (rubus glaucus benth.) genotypes," *Agronomía Colombiana*, vol. 38, no. 1, 2020.

[7] J. S. Gutiérrez Díaz, "Evaluación del efecto de dosis de n, p, k y ca sobre las propiedades químicas del suelo y la productividad de un cultivo de mora (rubus glaucus benth.)," *Facultad de Agronomía*, 2017.

[8] C. Moreno Guerrero, M. J. Andrade Cuvi, A. Terán Guerrero, A. Túqueres Ushca, and A. Concellón, "Efecto del uso combinado de radiación uv-c y atmósfera modificada sobre el tiempo de vida útil de mora de castilla (rubus glaucus) sin espinas," *Revista Iberoamericana de Tecnología Postcosecha*, vol. 17, 2016.

[9] G. A. Figueredo-Ávila and J. A. Ballesteros-Ricaurte, "Identificación del estado de madurez de las frutas con redes neuronales artificiales, una revisión," *Ciencia y Agricultura*, vol. 13, no. 1, pp. 117–132, 2016.

[10] J. S. Moreira, F. M. Palacios, M. F. Marmolejo, K. L. Hidalgo, and A. V. Chóez, "Aplicación móvil para control de maduración de frutillas utilizando algoritmo de procesamiento de imágenes de matlab," *International Journal of Innovation and Applied Studies*, vol. 29, no. 2, pp. 149–159, 2020.

[11] F. Martínez, H. Montiel, and F. Martínez, "Blueprints obtention by means of using digital image processing algorithms," *International Journal of Engineering and Technology*, vol. 10, no. 4, pp. 1129–1135, 2018.

[12] F. Martínez and H. Martínez, F. Montiel, "Hybrid free-obstacle path planning algorithm using image processing and geometric techniques," *ARPN Journal of Engineering and Applied Sciences*, vol. 14, no. 18, pp. 3135–3139, 2019.

[13] A. Rendón, "Design and evaluation of volume unit (vu) meter from operational amplifiers," *Tekhnê*, vol. 16, no. 2, pp. 31–40, 2019.

[14] F. Martínez, F. Martínez, and H. Montiel, "Minimalist 4-bit processor focused on processors theory teaching," *Indian Journal of Science and Technology*, vol. 10, no. 14, pp. 1–6, 2017.

[15] O. Espinosa, L. Castañeda, and F. Martínez, "Minimalist artificial eye for autonomous robots and path planning," in *International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2015)*, vol. 9375, no. 1, 2015, pp. 232–238.

[16] E. Jacinto, F. Martínez, and F. Martínez, "A cordic based configurable fixed-point design on fpga using minimal hardware," *Indian Journal of Science and Technology*, vol. 10, no. 24, pp. 1–5, 2017.

[17] L. Ríos and M. Sarrazola, "Prefactibilidad de crear una empresa productora y comercilizadora de pulpa de fruta de mora y fresa, para exportar a corea del sur," *ESUMER*, 2016.

[18] J. P. Castillo Quevedo and M. T. Riveros Gonzalez, "Costos de producción de mora por procesos en el municipio de san bernardo," *UdeC*, 2018.

[19] MinAgricultura. (2019) Subsector productivo de la mora. [Online]. Available: https://sioc.minagricultura.gov.co/Mora/Documentos/2019-12-30%%20Sectoriales.pdf

[20] M. F. Jaimes Flórez and J. F. Sanabria Vanegas, "Dinámica de la internacionalización de la mora en santander para su exportación a mercados potenciales," *USTA*, 2019.

[21] L. Isaza, Y. Zuluaga, and M. Marulanda, "Morphological, pathogenic and genetic diversity of botrytis cinerea pers. in blackberry cultivations in colombia," *Revista Brasileira de Fruticultura*, vol. 41, no. 6, pp. 1–6, 2019.

[22] L. Aguirre, L. Cubillos, M. Tarazona-Díaz, and L. Rodriguez, "Efecto del tratamiento y tiempo de almacenamiento sobre los compuestos funcionales de subproductos de mora y fresa," *Revista UDCA Actualidad & Divulgación Científica*, vol. 22, no. 1, 2019.

[23] A. S. Cardona, G. Franco, C. A. D. Diez, and G. E. M. Uribe, *Manual de campo para reconocimiento, monitoreo y manejo de las enfermedades de la mora (Rubus glaucus Benth)*. Corpoica Editorial, 2017.

[24] C. M. Grijalba Rativa, L. A. Calderón Medellín, and M. M. Pérez Trujillo, "Rendimiento y calidad de la fruta en mora de castilla (rubus glaucus benth), con y sin espinas, cultivada en campo abierto en cajicá (cundinamarca, colombia)," *Revista Facultad de Ciencias Básicas*, vol. 6, no. 1, 2010.

[25] Y. A. Zapata-Narváez and C. R. Beltrán-Acosta, "Evaluación de la eficacia de alternativas sostenibles para el control de enfermedades del cultivo de mora," in *Anais do Congresso Brasileiro de Fitossanidade*, vol. 5, no. 1, 2019.

[26] J. Alvarado, V. Aguilar, M. Molina, and A. Campoverde, "Clasificación de frutas basadas en redes neuronales convolucionales," *Polo del Conocimiento: Revista científico-profesional*, vol. 5, no. 1, pp. 3–22, 2020.

[27] H. Cecotti, A. Rivera, M. Farhadloo, and M. Pedroza, "Grape detection with convolutional neural networks," *Expert Systems with Applications*, vol. 159, no. 1, p. 113588, 2020.

[28] C. Borrego, "Evolución de compuestos de interés biológico en moras a lo largo de la maduración del fruto," Master's thesis, Universidad de Cádiz, 2018.

[29] J. Duarte, Y. Triviño, and F. Martínez, "Application of custom-made simulator for training in robotic navigation strategies," *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 5, pp. 1184–1196, 2021. [Online]. Available: http://www.jatit.org/volumes/Vol99No5/16Vol99No5.pdf

[30] V. Senén Cerdà, "Diseño, implementación y evaluación de una red neuronal convolucional de regresión en clasificación de naranjas," Ph.D. dissertation, Universidad Politécnica de Valencia, 2020.

A Comparative Analysis of Generative Neural Attention-based Service Chatbot

Sinarwati Mohamad Suhaili[1], Naomie Salim[2], Mohamad Nazim Jambli[3]
Pre-University, Kota Samarahan, Sarawak, Malaysia[1]
Faculty of Computing, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia[1,2]
UTM Big Data Centre, Ibnu Sina Institute for Scientific and Industrial Research,
Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia[2]
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak,
Kota Samarahan, Sarawak, Malaysia[3]

***Abstract*—Companies constantly rely on customer support to deliver pre-and post-sale services to their clients through websites, mobile devices or social media platforms such as Twitter. In assisting customers, companies employ virtual service agents (chatbots) to provide support via communication devices. The primary focus is to automate the generation of conversational chat between a computer and a human by constructing virtual service agents that can predict appropriate and automatic responses to customers' queries. This paper aims to present and implement a seq2seq-based learning task model based on encoder-decoder architectural solutions by training generative chatbots on customer support Twitter datasets. The model is based on deep Recurrent Neural Networks (RNNs) structures which are uni-directional and bi-directional encoder types of Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). The RNNs are augmented with an attention layer to focus on important information between input and output sequences. Word level embedding such as Word2Vec, GloVe, and FastText are employed as input to the model. Incorporating the base architecture, a comparative analysis is applied where baseline models are compared with and without the use of attention as well as different types of input embedding for each experiment. Bilingual Evaluation Understudy (BLEU) was employed to evaluate the model's performance. Results revealed that while biLSTM performs better with Glove, biGRU operates better with FastText. Thus, the finding significantly indicated that the attention-based, bi-directional RNNs (LSTM or GRU) model significantly outperformed baseline approaches in their BLEU score as a promising use in future works.**

***Keywords*—Sequence-to-sequence; encoder-decoder; service chatbot; attention-based encoder-decoder; Recurrent Neural Network (RNN); Long Short-Term Memory (LSTM); Gated Recurrent Unit (GRU); word embedding**

I. Introduction

Providing excellent customer service while engaging with their clients has become more pivotal than ever in today's digitally connected era. Companies engage with customers to assist them with pre-post sale items regularly upgraded due to technological advancements or the communication revolution. Over the years, face-to-face physical meetings and phone calls have been the two most dominant communication methods. Since the rise of the internet, various ways have evolved, from email to social media, installing the mobile application to fill out a form on a website, and eventually waiting for a follow-up. Recently, the increasing use of real-time messaging such as Twitter, Facebook Messenger, WhatsApp, Telegram, Slack, etc., has led to a fundamental transition in how people would prefer to connect with businesses. While most of these communication channels have common characteristics, including online chat, which initially relies only on humans to conduct mutual communication, the baton now is passed to virtual agents or assistants called chatbots. Chatbots, the trendy platform led by virtual assistants, function as customer service representatives who negotiate conversations with clients to improve the user experience and services.

Chatbots are the subsequent major advancement in conversational services, which allow some business companies to communicate through messaging systems, like Twitter and Facebook Messenger, based on artificial intelligence and machine learning. Chatbots can be defined as computer programs living in messenger applications and providing specific services via emulating an interaction with a human through text messaging or a virtual voice [1] [2]. Owing to the overwhelming prevalence of chatbots as messaging is the most commonly used customer assistance medium; therefore, there is a need for the company to invest in a chatbot to support serving their customers' needs as applying in the context of service chatbots. Consequently, companies can strengthen employees' productivity to serve more customers with other services.

Chatbots' primary purpose is to facilitate the conversation between machines and humans in natural language conversation; as in the human viewpoint, these interactions should resemble humans as closely as feasible. Consequently, achieving this has become a fundamental task, with numerous researchers seeking the optimal way for having a chatbot to behave like a human. An effective chatbot should be able to comprehend the user's message, retrieve appropriate information according to the given statement and respond accordingly so that the user perceives the conversation as human-like.

The existing chatbots work just on pattern matching inputs and then finding a scripted answer corresponding to the information presented. The downside to this technique is that it cannot lead to a completely satisfying conversation due to the limitation of discourse within a specific domain with a clear goal. To handle the user's input utterances, Eliza, PARRY, and ALICE, to name a few were among the first chatbots to employ rudimentary parsing, pattern matching, or keyword retrieval approaches. These techniques require handwritten rules to generate responses. Due to the domain-specific

nature of these practices, they were effective at preserving context. However, as the knowledge space expands and users' expectations upsurge, for instance, when engaging in chitchat as in [3], it then becomes difficult to predict users' intentions and considered as not cost-effective, as all the possible patterns must be built manually with a great deal of effort to have a large number of patterns for generating responses.

As artificial intelligence (AI), machine learning, and natural language processing (NLP) techniques advance, researchers and practitioners seek to use data-driven methods incorporating capabilities of deep learning techniques such as RNN, LSTM, and Sequence-to-Sequence (Seq2Seq) model in constructing chatbots automatically and minimizing hand-written rules chatbots techniques. Minimizing the hand-written rules requires the chatbots to be designed based on modular form. This modular form consists of several components: a natural language understanding (NLU) component that turns user text or speech input into semantic representation, an internal state update component that updates the conversation memory (dialogue state tracker), and dialogue policy are used to decide what the following system action will be (dialogue policy), and a natural language generation (NLG) component for producing a response to the user. Training each modular system component typically necessitates a considerable quantity of tagged dialogue data. In contrast to its end-to-end counterparts, the system is more interpretable and stable due to its modular design.

On the other hand, researchers and practitioners recently tried to implement the end-to-end approach utilizing the seq2seq learning task model based on the neural machine translation problem. This attempt is due to the fact that the end-to-end process needs less annotation, giving it an additional viable choice for commercial use cases [4]. In addition, the performance of each component in a modular system is not representative of the entire system because each element is optimized separately [5]. Nevertheless, its end-to-end design makes it uncontrollable [5].

Furthermore, the evolution of machine learning, AI and NLP techniques has encouraged the academics and developers to create chatbots that employ various design strategies. However, despite these advancements in design, chatbots still face several hurdles in comprehending incoming requests, interpreting them, providing acceptable replies, and sustaining a user dialogue. Therefore, academics and developers continue to improve chatbot development techniques to meet the demands of both consumers and service providers. Consequently, it is essential to find a new method to enhance the accuracy of the user utterance's understanding and chatbot's response in service chatbot application. Thus, the fundamental objective and contribution of the current paper aim to present and implement a seq2seq-based learning task model based on encoder-decoder architectural solutions by training generative chatbots on customer support Twitter datasets. These generative chatbots are important to predict automatically an appropriate and automatic response to customers' queries and extensively evaluate their effectiveness in a variety of circumstances under various baseline models, training hyperparameters, and architectures.

The remaining sections of the paper are organized as follows. In Sections II and III, reviews of related works and descriptions of the models are provided, respectively. The methodological approach is presented in Section IV. Section V contains the experimental study of the research, while Section VI includes the conclusion and recommendation for future work.

II. Related Work

Seq2seq learning task models have been implemented in numerous natural language processing tasks, such as chatbot, machine translation, question answering, text summarization, image captioning, sentiment analysis, etc. Initially, seq2seq learning comprising encoder and decoder(E2D) structure was introduced in [6] for Neural Machine Translation (NMT). With the support of gate mechanisms such as LSTM [7] and GRU [8], the problem of vanishing or explosion can be controlled, enabling the model to obtain far longer sentences.

To better capture the dependencies in utterance, bidirectional and reverse order practices are commonly used to design the seq2seq models [9] [10]. Yet, at the same time, this approach also has a fixed-length vector (context vector) issue. This issue arises in the decoding process because the source sentences will compress the input regardless of the length vector this neural network needs as a context vector, especially when the source sentence is long [8]. Indirectly, this process leads to incorrect responses, as each word in the answer may have a close relationship with various sections of the words in the request.

A study was done in [11] and [12] combated the problem by adding an attention mechanism layer and integrating it into the decoder by repeatedly reading the representation of a source sentence, which remains fixed after being generated by the encoder. Hence, the model is able to search for relevant parts to predict a targeted word and attain cutting-edge machine translation performance. Inspired by the great commission in machine translation, the researchers and developers attempt to apply this technique to other tasks, including chatbots. For instance, in [13], the author also used seq2seq learning by comparing the performance of chatbot-generated responses between LSTM, GRU, and Convolution Neural Network (CNN). Unlike the above approaches, our work integrates an additive attention model to align the relevancy of input and output for improving question-answer. In addition, a study in [14] also employed attention mechanisms to the encoder-decoder architecture in enhancing question-answer relevance.

In [15], the authors implemented an attention-based with encoder-decoder neural architecture with the knowledge graph, and the corpus joins embedding as input in a task-oriented based chatbot. In contrast to [16], the authors added information regarding the conversation history and external knowledge collected from the search engine to enhance the seq2seq chatbot. In [17], the authors allocated labeled data to hierarchical categories using the attention-based Seq2Seq model. In the research, when answer predictions were inconsistent, a slot-filling method was used to determine which questions needed to be asked in order to make correct predictions.

All the above-related works focus less on examining the effects of complementary mechanisms in deep learning, such as batch size, lstm size or types of embedding in different seq2seq chatbot architecture. Therefore, this work attempts to

focus more on investigating such effects through experimental study as presented in Section V.

III. Model Description

This section describes and explores our model architecture based on the word representation model, seq2seq RNNs (LSTM, GRU, biLSTM, biGRU), and an attention mechanism.

A. Word Representation Model

For a computer to comprehend the meaning of the words and sentences, the text data must be converted to a numerical format. Embedding (encoding or vectorizing) is the term used for this concept. Character embedding, word embedding, and phrase embedding are only a few examples of many other types of embedding. Among various types of embedding, the word embedding is most commonly employed [18]. Word embedding is a way to model language that maps words to vectors of real numbers. It encodes words or sentences with several dimensions in vector space. The embedding layer can be initialized using pre-trained word vectors such as Word2vec, Glove, or FastText as implemented in this research word. A detailed description of the word embedding models is presented in the following subsection:

1) Word2Vec: Word2vec is a predictive embedding technique that uses the low dimensionality of word vectors to learn fine word vectors from massive data sets containing billions of words. There are two main architectures of Word2Vec for producing a distributed word representation, namely:

- Continuous bag-of-words (CBOW) model
  - This architecture is based on the language model of a feed-forward neural network [19]. It seeks to anticipate the current word based on the surrounding context by minimizing the loss function
- Skip-Gram model
  - Unlike the CBOW model, this model is aimed to predict surrounding words given the current word.

2) Glove: The Global Vectors for Word Representation (GloVe) enhances the Word2vec approach proposed in [20] at Stanford in 2014 for effectively learning word vectors. Conceptually, the Word2vec approach only considered local contexts but did not utilize a global context. Previously, the conventional vector space model representation of words was built using matrix factorization techniques such as Latent Semantic Analysis (LSA), which gave a better result than global text statistics. This model is also called a count-based model. Count-based models learn their vectors by reducing the co-occurrence counts matrix's dimensionality. However, this technique does not give a promising result as a learned method or predictive model such as Word2Vec captures the meaning and performs an arithmetic operation that can pose semantic or syntactic relationship of words, for example, $king - man + woman \rightarrow queen$. Thus, by merging the global matrix factorization and local context window approaches with the help of a bilinear regression model, GloVe indirectly benefits from both techniques.

3) FastText: FastText, made available by Facebook, Inc, is one of the contributions to prediction-based word embedding models. The usually cited work for this model is from [21] and [22]. The motivation behind the FastText model is because of the shortcoming of word embedding models that disregard the word's morphology and learn a distinct vector for each word.

The improvement is made based on the Skip-Gram model introduced in [23], wherein each word is represented as a bag of character n-grams. Each word is mapped to a set of n-grams, and the skip-gram model is modified to regard each word vector as the sum of its n-grams, which is based on the assumption that similar groups of letters express identical meanings.

Since word vectors are composed of known n-grams, hence can also be computed on unknown words. Consequently, even a word not in the vocabulary is assigned a vector based on its subword units. This unknown term is even more essential for inflected languages, as some inflected forms of words are uncommon and may not even be present in the training set. Training the FastText embedding is faster than the majority of other options due to their simplicity and efficient implementation.

B. Sequence-to-Sequence (seq2seq) Learning Task Model

Seq2seq learning task was initially proposed in machine translation for training models to map the sequence of input between one domain (e.g., German sentences) to output sequences in a different domain (e.g., the same sentence translated to English). Due to its promising result, many researchers investigated and worked by adopting this technique in various tasks such as image captioning, text summarization, and chatbot (question and answering task). The bot generates a natural language response as an output sequence given a natural language question as an input sequence.

The most typical architecture for constructing a seq2seq learning problem is using the encoder-decoder architecture. This architecture in the seq2seq learning task manifested in three parts: encoder, context vector (final hidden/internal state vector), and decoder, as the name implies. The encoder attempts to convey the meaning of the input sentence by encoding it into a fixed-size hidden representation. This hidden representation is converted to output by a decoder. The fundamental structure of this model is based on two RNNs (or can be used as another type of RNNs such as LSTM/GRU for better performance) [6]. Encoding the input into a vector representation employs one RNN as the encoder that captures the context and essential information of the input sequence. On the other hand, the other RNN (as a decoder) will then take this vector as input and use it to generate the output sequence. The basic architecture of the seq2seq model in training mode is illustrated in Fig. 1.

Based on this figure, let $x = \{x_1, x_2, x_3, ..., .x_n\}$ represent the words contained in each input statement or utterance (where n being the statement's length) is mapped in the form of embedded representation (φ^{x^n}) and passed to a variant type of RNN models such as LSTM or GRU. The embedded representation can be either pre-trained embedding such as Word2Vec, Glove, and FastText or jointly trained during model training to convert words into dense vectors, as mentioned

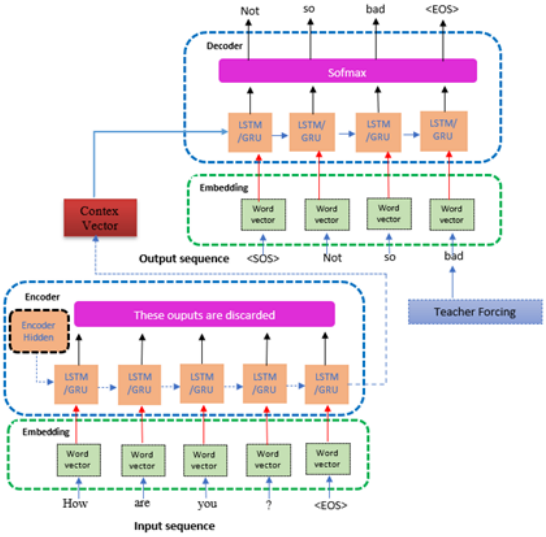


Fig. 1. Basic Encoder-Decoder Architecture While Training.

in the previous section. RNNs also take the first encoder hidden state as the input. RNNs take all these embeddings and sequentially give hidden representation and output vectors at each time step. Hence, it takes a word and a hidden state of the previous state as an input and provides one output and updated hidden state until it reaches the end of the input mark with a unique token known as <EOS>, the outputs at each time step of the encoder part are all discarded since outputs will be summarized by the context vector (C). This context vector contains information about all of the input items that enable the decoder to predict accurately. Equations (1) and (2) illustrate the computation of hidden states and context vectors, respectively;

$$h_m = f_1(\varphi^{x^n}, h_{m-1}) \tag{1}$$

$$c = f_2(\{h_0, h_1, ...h_M\}) \tag{2}$$

where h denotes as the hidden state, c denotes the context vector constructed from the encoder hidden states and f_1, f_2 are nonlinear functions such as LSTM / GRU in this case.

The context vector also acts as the decoder's initial hidden state to pass information from the encoder to the decoder. Before passing to the decoder, this work considers uni-directional or bi-directional encoders, wherein bidirectional encoder; there will be one forward RNN and one reverse RNN. The processing of an input sequence occurs in both directions (forward and backward). The forward and backward hidden states are then combined before being transmitted to the decoder.

In the decoding phase, at the first timestep, the <SOS> token is given as input to RNNs along with the context vector. <SOS> marks the beginning of decoding, and it generates the first word of the chatbot response by looking at the context vector. This decoding first generated output of RNNs probably *"Not"*. For the next timestep, the *"Not"* will be given as the input along with the previous timestep hidden state. This step will provide an output as *"so"*. This output generation will continue until it reaches a unique token known as <EOS> is encountered. Considering the context vector as c, and all previously predicted output as $\{y_1, y_2, y_3, ...y_{t-1}\}$, the decoder has been trained to anticipate the following token y_t. This prediction is the maximum likelihood estimation of y_t. The prediction is given y, the output vector, and c, the context vector. Thus, the $p(y)$ is computed as in Equation (3):

$$p(y) = \prod_{t=1}^{T} p(y_t|y_1, y_2, y_3, ...y_{t-1}, x_t) \tag{3}$$

and produces a token with a conditional probability for each timestep t through the following Equation (4):

$$p(y_t|y_1, y_2, y_3, ...y_{t-1}, x_t) = g(y_{t-1}, s_t, c_t) \tag{4}$$

where $g(.)$ is a softmax function and s_t is the decoder's hidden state at the timestep t which can be computed as in the Equation (5) as follows:

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \tag{5}$$

1) Long Short Term Memory (LSTM): LSTM was developed as a short-term memory solution and initially proposed in [7]. LSTMs are a type of RNN that uses specific hidden states to manage long-term dependencies better while memorizing inputs over time [7]. The difference from standard RNN is how the hidden state is calculated within LSTM cells. LSTMs architecture has an internal cell state that acts as a transport highway that can carry and filter the information in a sequence by adding or removing it. The adding and removing information state is controlled by a structure known as gates.

Gates have sigmoid activations identical to the tanh activation, whereas rather than squishing values from -1 to 1, it squishes values from 0 to 1. This value is significant when updating or removing the data as every value multiplied by 0 is 0, allowing the values to be vanished or be 'forgotten'. On the other hand, every value multiplied by one will result in the same value, remaining unchanged or 'preserved'. Subsequently, the network learns which information is irrelevant and may be removed and which information should be retained with this activation function. Three more gates are generally added in contrast to standard RNNs cells, namely, forget, input, and output gates, as shown in Fig. 2.

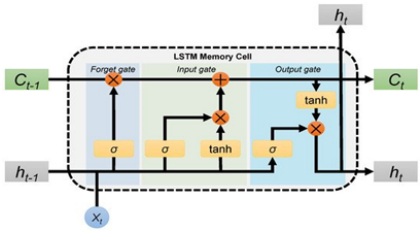


Fig. 2. The Long Short-Term Memory Introduced in [7].

2) Gated Recurrent Units (GRU): GRUs is a more recent generation of RNN cells compared to an LSTM introduced in [8]. The author has introduced two gates: the update gate and the reset gate, as depicted in Fig. 3. The update gate combines the forget and input gates in GRUs and operates similarly to LSTMs, controlling what information to discard and add. On the other hand, the reset gate is a different gate used to

specify the amount of prior information that can be discarded. In GRUs, it also merges the hidden state and cell state, and thus, the output gate is no longer needed. Therefore, this model is simpler while gaining more popularity than regular LSTM, with fewer parameters and faster training than LSTM [8].

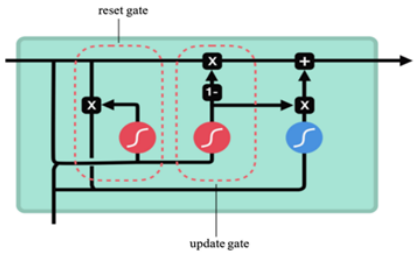


Fig. 3. Types of Gates in GRU Cell Introduced in [8].

3) Bi-Directional RNNs Cell: The previous section discusses the implementation of uni-directional RNNs, which means it runs in a single direction. This research attempts to employ bidirectional RNNs (both for LSTM and GRU) to improve model performance by incorporating past and future context information. This bidirectional means each layer has two RNNs: one running in the forward direction of the sequence (from left to right) and another running in the backward direction (from right to left) to capture dependencies in two contexts [24]. The resulting forward and backward outputs are concatenated before being passed on to the next layer, as shown in Fig. 4. The encoded representation of each word now has the information of the reverse and the future words of the particular word to predict output better.

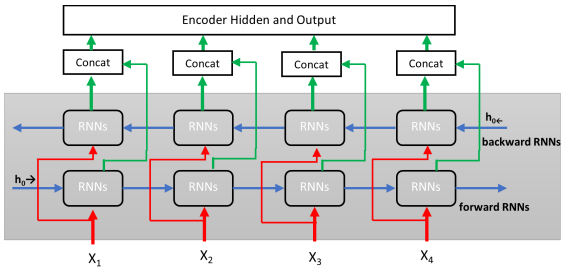


Fig. 4. The Bi-Long Short-Term Memory (biLSTM) Process Captures Sequential Features. The Last Hidden Layer of the LSTM is Extracted as Features Representing the Text.

4) Neural Attention Mechanism: The previous section describes the seq2seq model based on RNNs. These RNN structures utilize the temporal dynamics of the input data to generate sequential output data. However, the output created at a particular timestep and the input sequence used to obtain that result may or may not be relevant remains uncertain. Moreover, in the seq2seq model using RNNs, all the intermediate states of the encoder output will be discarded, and only its final states (context vector) will be used to initialize the decoder. This technique incorporates well with short or medium sequences; however, as the lengths of the sequence grow, a single vector becomes congested and more challenging to analyze long sequences into a single vector.

Meanwhile, RNNs sequentially process tokens while preserving a state vector representing the data observed after each token. The information from the inputs can be arbitrarily propagated in the sequence through the continuous encoding of the data. Due to the vanishing gradient problem, the model's state towards the end of a long sentence typically does not have information about earlier tokens. As a result, the process does not perform as expected. Long sequences benefit from LSTM, reducing disappearing and exploding gradient effects, albeit not entirely eliminated. Furthermore, RNN architectures may be unable to handle increasingly complex feature representations in order to produce reliable outputs.

The aforementioned issue was resolved with the introduction of attention mechanisms introduced in [11] and [12]. Attention processes enable a model to directly examine the condition of an earlier point in the sentence and derive conclusions from it. The attention layer has access to all previous states. It can weigh them according to some learned measure of relevance to the present token, allowing it to provide more precise information on distant relevant tokens, as illustrated in Fig. 5. It decides which source elements are the most important at each decoder step. The encoder does not need to condense the whole source into a single vector in this case; instead, it provides token representations for all of the source data (for example, all RNN states instead of the last one). In addition, the key concept behind attention is not to throw away these intermediate encoder states but to make use of all the states to create the context vectors that the decoder uses to produce the output sequence through attention weight. The attention weight is computed to decide which part of the input was relevant and subsequently determine the output.

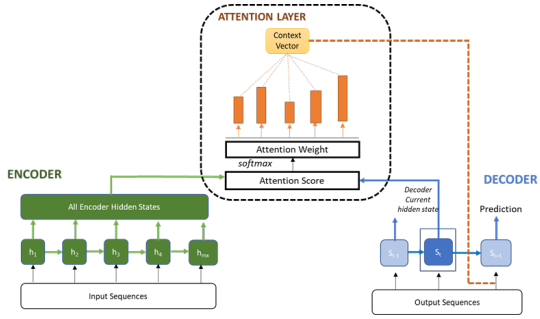


Fig. 5. Attention-Based E2D.

In the attention process, the relevance of each word in the input sequence will be determined for each output cell. For each y_t in the output y, it is influenced by the context vector c_t (source context for decoder step t) are used in an information filter for all hidden states $h = \{h_1, h_2, h_3, ..., h_{m_x}\}$ of the encoder, which can be computed as in the following Equations (6), (7) and (8):

$$c_t = \sum_{i=1}^{m_x} \alpha_{t_i} h_i \tag{6}$$

Where α_{t_i} is calculated by

$$\alpha_{t_i} = \frac{exp(e_{ti})}{\sum_{j=i}^{m_x} exp(e_{tj})} \tag{7}$$

Where $e_{ti} = align(s_{t-1}, h_i)$ refers to the additive score function that considers

$$e_{ti} = V_a^T tanh(W_a s_{t-1} + U_a h_i) \quad (8)$$

Where α_{t_i} indicates the attention weights that the model has learned, W_a ,U_a and V_a, implies another weight parameter for the model to learn. The align is an alignment model for evaluating the relationship between the input of position i and the output of the position t.

IV. Methodological Approach

This section gives an overview of the current research methodological approach to previously discussed implemented models. Fig. 6 depicts the methodology steps involved in this work as follows:

- *Dataset Preparation* – The dataset is collected using publicly available datasets from free online websites which we examine the data and explore the dataset using some fundamental Exploratory Data Analysis (EDA).
- *Data Prepossessing* – Load the text, perform pre-processing or data cleaning, and do a train-test split. In this phase, we build questions-answer pair. Append <START> and <END> to all the answers. Create a Tokenizer and load the whole vocabulary into it with the help of embedding techniques for feature extraction.
- *Modeling* – Define the model. We implement an encoder-decoder architecture-based seq2seq learning task model with and without attention to a different variant of RNN cells
- *Training and tuning* – The model will be trained and tuned with the help of various hyperparameter optimization and regularization techniques to overcome overfitting during training. We aim to minimize the objective function during training by reducing the loss.
- *Results* – The trained model will be evaluated using a valid/test set through BLUE score and validation loss during training based on predicting answers.

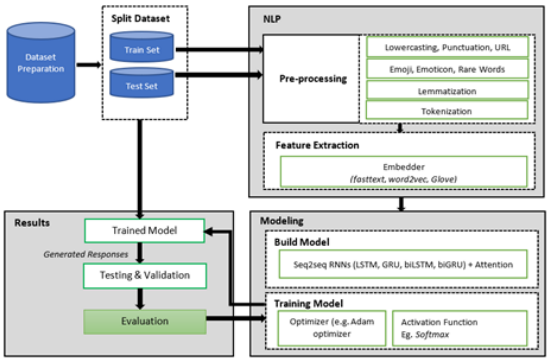


Fig. 6. Illustrates the Methodology Steps.

V. Experimental Study

This section presents different experiments conducted to study the functionality of various models employed (as mentioned in the previous section) and thoroughly examine their performance. The experimental results are compared based on the effect of several variables such as the encoder types, adding an attention layer, variant of embedding, the number of hidden sizes, and batch sizes. The dataset used in the experiment, experiment settings, evaluation methods, and qualitative analysis of experimental results are described in the following subsections. before

A. Datasets

The experiment trains and evaluates the models using the *"Customer Support on a Twitter (CST)"* dataset from Kaggle[1]. The CST dataset was collected in 2017 with a huge, innovative corpus of tweets and replies for the advancement of NLU and conversational models, along with research into modern customer service techniques and impact. It consists of 2,811,774 tweets and replies, with 1,537,843 (54.69%) tweets generated from consumers and 1,273,931 (45.31%) generated from customer supports agent. Among these 1.5 million customer tweets, about 1.27 million received replies from customer support agents, and 0.23 million otherwise. One of the main reasons to use this dataset is that it contains real-life conversations between customers and customer support agents with natural responses from support agents for accurately explaining problems and solutions. Moreover, it is practical since it permits a relatively small message size restriction for recurrent networks.

While conducting an exploratory analysis of the dataset, it can be observed that there are 108 customer support brands represented in the dataset, and 597075 consumers' requests are answered. The top 20 customer support replies related to the company brand are depicted in Fig. 7.

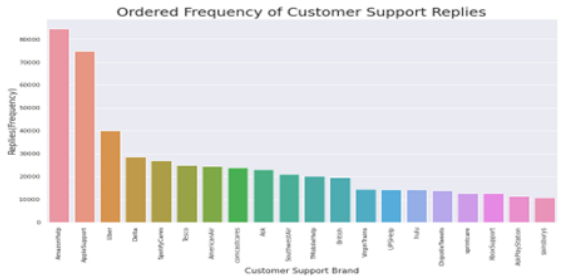


Fig. 7. The Top 20 Customer Service Replies by Brand.

In addition to this observation, it was identified that Amazon's customer service responded to a large number of inquiries, followed by Apple and Uber. There are a lot of companies in the dataset that have minimal responses or had no responses at all.

As shown in Table I, the information in the dataset must be restructured to create a conversational dataset between consumer and customer support agents suited for the current

[1]https://www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter

TABLE I. DATASET FEATURES DESCRIPTION

Features	Description	Datatypes
tweet_id	The unique ID for this tweet	Int64
author_id	The unique ID for this tweet author (anonymized for non-company users).	Object
Inbound	Whether or not the tweet was sent (inbound) to a company on Twitter. This feature is useful when re-organizing data for training conversational models.	Bool
created_at	Date and time when the tweet was sent.	Object
Text	The text content of the tweet.	Object
response_tweet_id	IDs of tweets that are responses to this tweet.	Object
in_response_to_tweet_id	IDs of the tweet is in response to, if any.	Float64

study. In the restructuring process, the dataset features must be filtered by selecting only inbound tweets that are not retweeted. Then, apply the *"in_response_to_tweet_id"* and *"tweet_id"* features to associate each tweet with the relevant reply based on the inbound feature, excluding instances where response tweets are not from a company. As the data is still in unstructured text data, additional preprocessing is required to eliminate unnecessary features such as emoji and emoticons, lower casting, non-English tweets, etc. The new dataset has 794,299 rows and six columns consisting of *'author_id_x'*, *'created_at_x'*, *'text_x*, *author_id_y'*, *'created_at_y'*, and *'text_y'*, where *x* and *y* are represented as a question from consumers and answered by customer service agent respectively.

Training and validation are conducted independently on 75% and 35% of the entire dataset. The model is termed accurate if the predicted response matches the ground-truth answer. This current research incorporates the Bilingual Evaluation Understudy (BLEU) score function to evaluate the performance models.

B. Experimental Settings

The study compares the attention-based approach to related baseline models that do not employ attention mechanisms to evaluate how well the models work. These models are implemented in a python-dependent package on a deep neural network framework called TensorFlow [25] and Keras[2]. We trained models on a GPU with 3082 CUDA cores and a VRAM of 12GB. The model is trained for 500 epochs (a high value is set since the study employs the early stopping technique) and tested on a batch size of 64, 128, 256, and 304 (the number can be divisible by 8). While the hidden size of LSTM and GRU is tested on 100, 200, and 300 units. For the optimization, the study uses the Adam optimizer with a learning rate of 0.003 [26]. A gradient clipping of 50.0 is implemented to combat the 'exploding gradient' problem, preventing the gradients from expanding exponentially and causing the cost function to either overflow (with undefined values) or overshoot cliffs. All the weights and biases are initialized using Xavier and glorot uniform distribution [27].

This study uses 300-dimensional pre-trained word embeddings for Fasttext[3], GloVe[4], and Word2Vec[5]. An early stopping technique with patience four is adopted to combat overfitting. Table II shows the hyperparameters and their corresponding ranges for training the models. It also presented the best-performing hyperparameter for each of the models.

TABLE II. HYPERPARAMETER SETTING

Parameter	Range	Final Setting
Max input Length	39	39
Word embedding	FastText/Glove/Word2vec	FastText/Glove
Embedding size	300	300
Encoder types	Unidirectional/Bidirectional	Bidirectional
Learning Rate	0.003	0.003

C. Performance Evaluation Metrics

Following [13] and [15], this current study adopts BLEU as suggested in [28] best to evaluate the performance of our model. According to the definition provided in [28], BLEU evaluates the co-occurrences of n-grams in the reference human translation and recommended answers. It computes the n-gram precision for the entire dataset, compounded by a brevity penalty to penalize brief translations. The more closely a machine translation resembles a professional human translation, the better.

The BLEU score compares the chatbot-generated output text (hypothesis) to a human-generated response text (reference). It specifies how many n-grams from the output text are included in the reference. The BLUE score can take on any value within the interval [0, 1] and is technically defined as Equation (9).

$$BLEU = BP \times \left[\prod_{n=1}^{N} precision_n\right]^{1/N} \tag{9}$$

Where N is the maximum n-gram number $n = [1, N]$ (N=4 in our evaluations). BP (brevity penalty) and $precision_n$ are defined with Equation (10) and (11), respectively.

$$BP = min(1, exp(1 - \frac{ref_{lenght}}{out_{lenght}})) \tag{10}$$

Where ref_{lenght} is reference length, and out_{lenght} is the chatbot output length

$$precision_n = \frac{\sum_n min(m^n_{out}, m^n_{ref})}{\sum_{n'} m^{n'}_{out}} \tag{11}$$

Where m^n_{out} is the number of n-grams matching the reference in the chatbot output, m^n_{ref} denote as the number of n-grams in the reference, and $\sum_{n'} m^{n'}_{out}$ implies the total number of n-grams in the output of the chatbot.

The BLEU scores were computed using the blue score module from the translated package on the nltk[5] platform, which was built in Python.

[2]www.keras.io.

[3]http:fasttext

[4]http://nlp.stanford.edu/projects/glove/

[5]https://code.google.com/archive/p/word2vec/

[5]Natural Language Toolkit: Available online: https://www.nltk.org/

D. Results and Comparison

This section elaborates on the experimental results from our model on the mentioned dataset. The experiment assessed the performance of the different models based on encoder types and various types of embedding with varying parameters by comparing models to baseline RNNs such as LSTM and GRU. After multiple experiments, the study concluded the result of the promising hyperparameter and settings as presented in Table II. One of the models uses a pre-trained 300-dimensional word embedding as a hidden size tested on varying numbers of 100, 200 300, where the hidden size of 300 gives a promising result of a pre-trained 300-dimensional word embedding.

The experimental results on all the models indicated that the bidirectional encoder type attention-based models achieved promising performance and outperformed the neural networks that do not use an attention mechanism as a baseline. The different hyperparameters are used to test each model, and the promising performing hyperparameters for each model are shown in Table II. The impact of some of these hyperparameters will be highlighted, and an example of generated response based on several models is presented in the following subsections.

1) Effect of Encoder Type in E2D Architecture: This section evaluates the performance of the two distinct encoder types used in these models. In comparison to a uni-directional encoder, the model performance with bi-directional encoder types yields promising results, as shown in Fig. 8. This situation might happen because the bi-directional encoder works by preserving the information from both sentence directions, allowing the network to predict the next word better as it can understand the sentence context more. Thus, the bi-directional encoder is promising to be used for further analysis in this study.

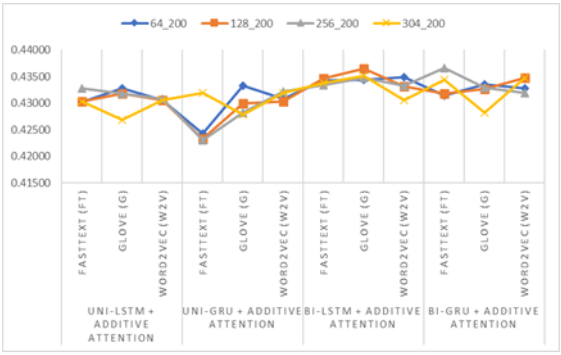


Fig. 8. E2D Network of RNNs with Different Encoder Types and Embeddings.

2) Effect of Attention Layer in E2D Architectures: This section uses hyperparameters from Table II to evaluate the RNNs E2D model with and without attention. As depicted in Fig. 9, the model's performance with attention vastly outperforms the RNNs E2D model without attention on this dataset. Therefore, the attention-based models improve the predictive performance of the Seq2Seq chatbot models. Furthermore, in comparing the performance of the RNNs types, the LSTM gives promising results in almost all models, as shown in Fig. 8. However, the GRU model can be an option if computational time is considered since GRU trains faster than LSTM, and the result for this dataset is not much different than LSTM.

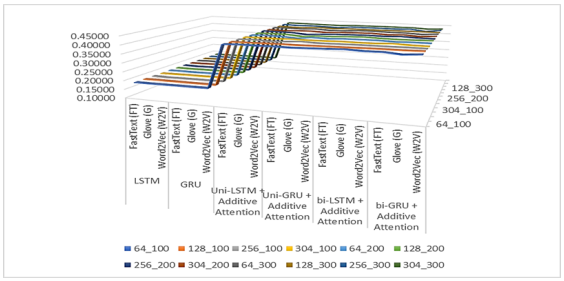


Fig. 9. Encoder-Decoder Network Model with and without Attention.

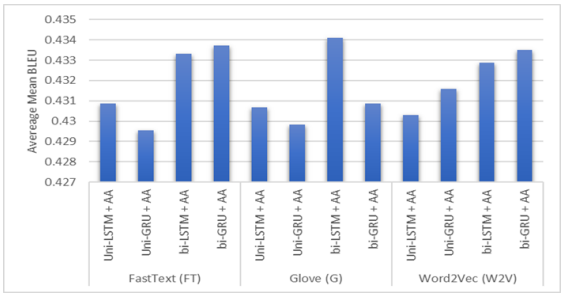


Fig. 10. Different Embedding Types in Encoder-Decoder Architecture.

3) Effect of Different Embedding Types in Encoder-Decoder Architectures: The performance of the RNNs models with varying types of embedding (FastText, Glove, and Word2Vec) is analyzed by finding the BLEU score among different variants of the models. As shown in Fig. 10, FastText and Glove embedding types give better performance for this type of chatbot dataset. Moreover, these embedding types perform better when integrating with a bi-directional encoder with LSTM (for Glove embedding) or GRU (for FastText).

4) Effect of Hidden Sizes in E2D Architectures: The performance of RNNs encoder-decoder models with different hidden sizes/units is evaluated in this section. The hidden size parameter is tested on 100, 200, and 300 units with a fixed embedding size of 300, as in Table II. Most models produce better results for 200 or 300 units, where 42% of each out of total models occur, as presented in Table III. Since a fixed embedding size of 300 is used, the good hidden size is preferable to be an equal size with the embedding size as indicated in this experiment result for this dataset.

5) Effect of Batch Sizes in E2D Architectures: The different batch sizes evidenced in this experiment were $B = [64, 128, 256, 304]$ with hyperparameter setting as in Table II. As illustrated in Fig. 10, the smaller batch size of 64 produces better results amongst models. Furthermore, the batch size of 64 with 300 hidden sizes produced more promising results in this experiment dataset, indicating best used for further analysis in this research. The smaller batch sizes correlate to stability during the training and are better for accuracy. On the other hand, the larger batch size is better for computational speed. Nevertheless, batch size should be adequate so that the data would fit into memory. Due to the limitation of memory

space, the experiment can be done with a maximum batch size of 304.

VI. Qualitative Analysis

To qualitatively comprehend the model's performance gained with a different experiment configuration, as shown in Table II, the study prepares the outputs responses for a specific customer inquiry. The customer's query may be based on an emotive or informative question. For the informative type customer query instances, *"When is delta is last scheduled flight?"*, predicted responses are given in Table III. On the other hand, Table IV shows an example of an emotional types question.

As observed, most RNNs models' response prediction can provide general and reasonable answers for customer queries. However, both types of questions can predict better responses by implementing bi-directional RNNs based on the actual or target response. Based on the emotional type query, the bidirectional with attention models can predict the specific word, the *"account"*, rather than *"details"* (a general response but acceptable) as referred to in the actual reaction. Moreover, the models are more emotional in responding to customer queries than humans for both types of questions. Conversely, the informative queries and asking for particular information requests are difficult to formulate, and the resulting responses are less pertinent to the question. However, by looking at the answers of various models, it was discovered that bidirectional LSTM with attention generates better logical statements, and the response appears to be indistinguishable from the actual one.

TABLE III. Example of Generated Response for Informative Type Question (64_300 FastText)

Customer Questions	When is delta is last scheduled flight?
Seq2seq E2D LSTM	hi there i am sorry to hear that you are having a problem with the app please do your confirmation number i will be happy to help tab
Seq2seq E2D GRU	i am sorry to hear that you are having trouble with your flight please do your confirmation number so i can take a look at your flight tow
Seq2seq E2D LSTM Unidirectional + Attention	hi there we are not showing any announcements on this flight if you have any other questions please let us know by
Seq2seq E2D GRU Unidirectional + Attention	i am sorry for the delay in your flight i am showing that you are not able to get a gate to clear your flight tow
Seq2seq E2D LSTM bidirectional + Attention	i am sorry to hear that you are not able to get through to us please do me your confirmation number and i will gladly look into it for you ¡end¿
Seq2seq E2D GRU bidirectional + Attention	hi there we are not sure what you mean we are here to help please do us your confirmation number so we can follow up ¡end¿
Actual	**certainly as soon as we get a response from the appropriate team that has the info we will be sure to tweet you**

VII. Conclusion and Future Work

As a conclusion, we attempt to construct a generative service chatbot to automatically respond to customers' queries for assisting companies in having a 24-hour support system. To construct this system, seq2seq learning task models based on encoder-decoder architectures (with and without attention as the baseline) are deployed. This study further investigates these models through different RNNs structures (LSTM and GRU), encoder types (uni-directional and bidirectional), and different embedding types (FastText, Glove, and Word2Vec) and tested with varying parameters of training.

TABLE IV. Example of Generated Response for Emotional Type Question (64_300 FastText)

Customer Questions	you have lied gave me the run around have the worst customer service ever so disappointed i cant work without internet
Seq2seq E2D LSTM	i am sorry to hear that you are having trouble with your internet service i would be happy to help please do me the details of your issue
Seq2seq E2D GRU	i am sorry to hear that you are having trouble with your internet service i would be happy to help please do me the details of your issue
Seq2seq E2D LSTM Unidirectional + Attention	i apologize for the issues you are having with the services have your concerns been addressed if not i will be
Seq2seq E2D GRU Unidirectional + Attention	i am sorry for the poor experience can you please do pm the full service address and name on the account as we
Seq2seq E2D LSTM bidirectional + Attention	i am sorry to hear about the poor experience can you please do the full service address and name on the account as we assist
Seq2seq E2D GRU bidirectional + Attention	i am sorry to hear that you are having issues with your internet services please do your account details to help
Actual	**it is unfortunate you are having trouble to better assist please do me the account number thank you ami**

Based on the experimental results, the bi-directional RNNs (LSTM and GRU) attention mechanisms produced promising results for further work in a chatbot. Moreover, there are not many effects typed of embedding as well as batch size and hidden sizes for this dataset since the result is not significantly different. However, the stability of the result is obtained from batch size 64, with 300 hidden sizes (as the same value with embedding size) combining with FastText or Glove can be opted for further work. Additionally, based on the findings for this dataset, it is proven that while biLSTM performs better with Glove, biGRU operates better with FastText. Overall, adding an attention layer improved the BLEU score compared to the baseline models.

The significant of findings from this research work is that the proposed attention mechanism has demonstrated its ability to meet the aforementioned requirements. Therefore, this proposed work provides an extension of the existing technique in developing chatbots. The attempt is made through end-to-end approaches by using seq2seq learning task model adoption from neural machine translation. A potential direction for future research can be explored and examine the variant of attention mechanism based on different scoring functions in comparing to the current experimental results and findings. In addition, the experiment also can be investigated in other deep learning architectures such as generative adversarial neural network and training parameters, including varying learning rates, dropout rates, optimizer, and activation functions.

Acknowledgments

This research is partly funded by Ministry of Higher Education Malaysia under grant R.J130000.7851.4L942. The authors would also like to thank the Universiti Malaysia Sarawak (UNIMAS) and Universiti Teknologi Malaysia (UTM) for providing the resources used in this research work.

REFERENCES

[1] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology," *Artificial Intelligence Applications and Innovations*, vol. 584, pp. 373 – 383, 2020.

[2] M. F. McTear, "The rise of the conversational interface: A new kid on the block?" in *Future and Emerging Trends in Language Technology. Machine Learning and Big Data - Second International Workshop, FETLT 2016, Seville, Spain, November 30 - December 2, 2016, Revised Selected Papers*, ser. Lecture Notes in Computer Science, J. F. Quesada, F. Martín-Mateos, and T. López-Soto, Eds., vol. 10341. Springer, 2016, pp. 38–49. [Online]. Available: https://doi.org/10.1007/978-3-319-69365-1_3

[3] Z. Yan, N. Duan, J. Bao, P. Chen, M. Zhou, and Z. Li, "Response selection from unstructured documents for human-computer conversation systems," *Know.-Based Syst.*, vol. 142, no. C, p. 149–159, feb 2018. [Online]. Available: https://doi.org/10.1016/j.knosys.2017.11.033

[4] Z. Zhang, R. Takanobu, M. Huang, and X. Zhu, "Recent advances and challenges in task-oriented dialog system," *CoRR*, vol. abs/2003.07490, 2020. [Online]. Available: https://arxiv.org/abs/2003.07490

[5] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational ai," 2018, cite arxiv:1809.08267Comment: Foundations and Trends in Information Retrieval (95 pages). [Online]. Available: http://arxiv.org/abs/1809.08267

[6] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112. [Online]. Available: https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, A. Moschitti, B. Pang, and W. Daelemans, Eds. ACL, 2014, pp. 1724–1734. [Online]. Available: https://doi.org/10.3115/v1/d14-1179

[9] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, p. 3776–3783.

[10] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A new chatbot for customer service on social media," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 3506–3510. [Online]. Available: https://doi.org/10.1145/3025453.3025496

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, Eds. The Association for Computational Linguistics, 2015, pp. 1412–1421. [Online]. Available: https://doi.org/10.18653/v1/d15-1166

[13] M. Aleedy, H. Shaiba, and M. Bezbradica, "Generating and analyzing chatbot responses using natural language processing," *International Journal of Advanced Computer Science and Applications*, 2019.

[14] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma, "Topic aware neural response generation." in *AAAI*, S. P. Singh and S. Markovitch, Eds. AAAI Press, 2017, pp. 3351–3357. [Online]. Available: http://dblp.uni-trier.de/db/conf/aaai/aaai2017.html{#}XingWWLHZM17

[15] F. Kassawat, D. Chaudhuri, and J. Lehmann, "Incorporating joint embeddings into goal-oriented dialogues with multi-task learning," in *European Semantic Web Conference*. Springer, 2019, pp. 225–239. [Online]. Available: https://jens-lehmann.org/files/2019/eswc_jointembedding_dialoguesystems.pdf

[16] Z. Wang, Z. Wang, Y. Long, J. Wang, Z. Xu, and B. Wang, "Enhancing generative conversational service agents with dialog history and external knowledge," *Comput. Speech Lang.*, vol. 54, pp. 71–85, 2019. [Online]. Available: https://doi.org/10.1016/j.csl.2018.09.003

[17] M. Patidar, P. Agarwal, L. Vig, and G. Shroff, "Automatic conversational helpdesk solution using seq2seq and slot-filling models," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Z. Broder, M. J. Zaki, K. S. Candan, A. Labrinidis, A. Schuster, and H. Wang, Eds. ACM, 2018, pp. 1967–1975. [Online]. Available: https://doi.org/10.1145/3269206.3272029

[18] S. Mohamad Suhaili, N. Salim, and M. N. Jambli, "Service chatbots: A systematic review," *Expert Syst. Appl.*, vol. 184, no. C, dec 2021. [Online]. Available: https://doi.org/10.1016/j.eswa.2021.115461

[19] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944966

[20] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[21] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," 2016, cite arxiv:1607.04606Comment: Accepted to TACL. The two first authors contributed equally. [Online]. Available: http://arxiv.org/abs/1607.04606

[22] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.zip: Compressing text classification models." *CoRR*, vol. abs/1612.03651, 2016. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1612.html{#}JoulinGBDJM16

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[24] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, "Text classification improved by integrating bidirectional lstm with two-dimensional max pooling." in *COLING*, N. Calzolari, Y. Matsumoto, and R. Prasad, Eds. ACL, 2016, pp. 3485–3495. [Online]. Available: http://dblp.uni-trier.de/db/conf/coling/coling2016.html{#}ZhouQZXBX16

[25] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[26] L. Mou and Z. Jin, *Tree-Based Convolutional Neural Networks: Principles and Applications*, 1st ed. Springer Publishing Company, Incorporated, 2018.

[27] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *AISTATS*, ser. JMLR Proceedings, Y. W. Teh and D. M. Titterington, Eds., vol. 9. JMLR.org, 2010, pp. 249–256. [Online]. Available: http://dblp.uni-trier.de/db/journals/jmlr/jmlrp9.html{#}GlorotB10

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

CapNet: An Encoder-Decoder based Neural Network Model for Automatic Bangla Image Caption Generation

Rashik Rahman[1]
Computer Science and Engineering
University of Asia Pacific,
Dhaka, Bangladesh

Hasan Murad[2]
Computer Science and Engineering
Chittagong University of
Engineering and Technology
Chattogram, Bangladesh

Nakiba Nuren Rahman[3]
Computer Science and Engineering
University of Asia Pacific,
Dhaka, Bangladesh

Aloke Kumar Saha[4]
Computer Science and Engineering
University of Asia Pacific,
Dhaka, Bangladesh

Shah Murtaza Rashid Al Masud[5]
Computer Science and Engineering
University of Asia Pacific,
Dhaka, Bangladesh

A S Zaforullah Momtaz[6]
Computer Science and Engineering
University of Asia Pacific,
Dhaka, Bangladesh

***Abstract*—Automatic caption generation from images has become an active research topic in the field of Computer Vision (CV) and Natural Language Processing (NLP). Machine generated image caption plays a vital role for the visually impaired people by converting the caption to speech to have a better understanding of their surrounding. Though significant amount of research has been conducted for automatic caption generation in other languages, far too little effort has been devoted to Bangla image caption generation. In this paper, we propose an encoder-decoder based model which takes an image as input and generates the corresponding Bangla caption as output. The encoder network consists of a pretrained image feature extractor called ResNet-50, while the decoder network consists of Bidirectional LSTMs for caption generation. The model has been trained and evaluated using a Bangla image captioning dataset named BanglaLekhaImageCaptions. The proposed model achieved a training accuracy of 91% and BLEU-1, BLEU-2, BLEU-3, BLEU-4 scores of 0.81, 0.67, 0.57, and 0.51 respectively. Moreover, a comparative study for different pretrained feature extractors such as VGG-16 and Xception is presented. Finally, the proposed model has been deployed on an embedded device for analysing the inference time and power consumption.**

***Keywords*—Bangla image caption generation; encoder-decoder; bidirectional long short term memory (LSTM); bangla natural language processing (NLP)**

I. Introduction

A picture is equivalent to million of stories. It is simple for people to narrate these stories, but challenging for machines to illustrate them. In the domain of intuitive systems, machine generated image captioning is an amalgamation of computer vision and NLP. Semantically and syntactically correct image caption generation is challenging for the machine compared to human beings. However, automatic caption generation from image content has a significant number of real life applications from the field of human machine interaction (HCI) to robotics.

According to World Health Organization (WHO), almost 2.2 billion people in the world have a near or distance vision impairment[1]. Automatic image caption generation plays a significant role for visually impaired people by converting the caption to speech to have a better understanding of their surroundings.

Automatic speech generation for the humanoid robot is a challenging task which involves generating caption by understanding the robot vision. Therefore, automatic image caption generation has considerable impact in the field of robotics. Content creation for social media platforms has become a professional sector which has created a large job sector for the young generation. However, content needs proper captioning before publishing in social media platforms. Therefore, providing automatic suggestions for image captioning is handy for content creators on social media platforms.

In recent years, image caption generation has become a relatively active field of research and therefore a significant number of research has been found in literature where most of the researchers focus on image caption generation in the English language [1], [2].

Though Bangla is the seventh largest language in the world with 215 million speakers globally[2], far too little effort has been devoted to Bangla image caption generation. Researchers have not addressed automatic image captioning in Bangla for a long period of time due to a lack of an enriched dataset. After development of required dataset, several researches have been conducted on Bangla caption generation from visual image [3], [4], [5], [6], [7].

However, the performance metrics given in the previous related work show that the quality of the generated Bangla image caption is not quite satisfactory. Therefore, there is a clear scope for further improvements in automatic Bangla image captioning. Moreover, we did not find any attempt to deploy the model on an embedded device.

[1] shorturl.at/hRWZ6
[2] $https://www.vistawide.com/languages/top_30_languages.htm$

The objective of this research work is to develop a image captioning model that can automatically generate Bangla caption with better performance compared to the models found in the previous related works. In addition to proposing an end-to-end system, the trained captioning model is deployed within an embedded device in order to evaluate the efficiency of the model.

We designed a model architecture using a deep learning based encoder-decoder model which takes an image as input and generates the corresponding Bangla caption as output. The encoder network consists of a pretrained image feature extractor while Bidirectional LSTMs are used in the decoder network for caption generation. We explore different pretrained image feature extractors such as VGG-16, Xception, and ResNet-50 for the encoder network. The model is trained and evaluated using a Bangla image captioning dataset named BanglaLekhaImageCaptions. We have achieved the best training results for the encoder-decoder model with the ResNet-50 pretrained feature extractor. The final training accuracy during convergence is 91% and the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores are 0.81, 0.67, 0.57, and 0.51, respectively, which is the state of the art result for automatic Bangla image captioning.

The major contributions of this paper are:

- Trained and validated the state of the art model for automatic Bangla image caption generation using an encoder-decoder based model architecture.
- Explored different pretrained image feature extractors such as VGG-16, Xception, and ResNet-50 for the encoder network and found the model with ResNet-50 provides the best BLEU score for Bangla image captioning.
- Finally, Deployed the proposed model on an embedded device for analysing the key performance metrics such as the inference time and power consumption.

The remaining sections of the paper are organized as follows: The literature review covered in the Section II. The Section III presents an overview of the dataset. Section IV provides a comprehensive breakdown of the proposed system. In Section V, all the experimental details during training and validation of our model are stated. In Section VI, findings and comparisons of this research is provided. The conclusion is located in the Section VII. The remainder consists of references.

II. Literature Review

In this section, we illustrate the evolution of research in the field of automatic caption generation for images. Moreover, the recent development of Bangla image captioning in literature is presented.

After conducting a rigorous literature review, we have found that there are two types of techniques based on traditional machine learning and deep learning to generate automatic image captioning [8], [9].

Early research on automatic image captioning utilises traditional machine learning techniques such as similar image retrieval based captioning [10] and template matching based image captioning [11]. However, the generated captions are limited by a predefined corpus with images or templates and their corresponding captions as labels. Therefore, traditional machine learning techniques fail to generate relevant image captions if the input image has significant differences from the predefined corpus.

The recent impressive progress in the fields of computer vision and NLP has paved the way to deploying deep learning techniques to generate image captions automatically. Image captioning involves vision encoding for a high-level understanding of image features and language decoding for caption generation using the features generated from vision encoding. The encoder-decoder based deep learning model is the most effective technique to address vision encoding and language decoding. In literature, vision encoder is designed using stacked Convolutional Neutral Network (CNN) [1], and graph-based network [2]. Moreover, various pre-trained feature extractors such as VGG-16, InceptionResnetV2, and Xception have been deployed for vision encoding [3], [12] The language decoder is implemented using variations of Recurrent Neural Networks (RNNs) such as LSTMs and GRUs [2]. In addition, self attention based transformer models are utilised to design the language decoder [13].

A considerable amount of effort has been devoted to developing automatic image captioning techniques in languages such as English [1], Chinese [14], Japanese [15], Arabic [16], Hindi [17] and German [18] where large datasets related to image captioning are already available.

Due to a lack of an enriched dataset, researchers have not addressed automatic image captioning in Bangla for a long period of time. However, after the development of the required dataset, Bangla image captioning has become an active research area among researchers. Table I presents an overview of related literature on Bangla image captioning with information on model architecture designed, dataset used during training and evaluation, and BLEU score as evaluation metrics to measure the quality of the generated caption by the model.

Rahman et al. [12] has developed the first Bangla image captioning dataset named BanglaLekhaImageCaptions. They have trained and evaluated an encoder-docoder model using their own dataset, where the encoder network utilises a pre-trained feature extractor called VGG-16, and the decoder network is designed using stacked LSTMs network. However, they have not calculated the BLEU score on their whole test dataset and have only reported the BLEU score for a few sample test images during evaluation, where the BLEU score is unsatisfactory.

Kamal et al. [6] have proposed a similar encoder-decoder model mentioned in [12] for Bangla image captioning where the encoder network consists of a VGG-16 pre-trained model and the decoder network consists of LSTMs network. Moreover, they have utilised the same BanglaLekhaImageCaptions dataset for training the model. However, they have evaluated the model by calculating the BLEU score for the test dataset, which was missing in [12]. The achieved BLEU-1 score for the model is 0.67 on the test dataset.

Jishan et al. [4] have proposed a hybrid encoder-decoder

TABLE I. AN OVERVIEW OF RECENT RESEARCH WORKS ON BANGLA IMAGE CAPTION GENERATION

Research	Year	Dataset	Modeling techniques	Performance
Humaira et al. [3]	2021	BanglaLekhaImageCaption	InceptionResnetV2 or Xception + BiLSTM or BiGRU	BLEU-1: 0.674, BLEU-2: 0.53, BLEU-3: 0.45, BLEU-4: 0.344
Khan et al. [5]	2021	BanglaLekhaImageCaption	1D CNN+ResNet-50	BLEU-1: 0.65, BLEU-2: 0.45, BLEU-3: 0.28, BLEU-4: 0.175
Palash et al. [7]	2021	BanglaLekhaImageCaption	ResNet-101+Attention mechanism+decoder	BLEU-1: 0.69, BLEU-2: 0.63, BLEU-3: 0.58
Kamal et al. [6]	2020	BanglaLekhaImageCaption	VGG-16+LSTM	BLEU-1: 0.67, BLEU-2: 0.44, BLEU-3: 0.32, BLEU-4: 0.24
Jishan et al. [4]	2020	BNLIT	CNN+BiLSTM	BLEU-1: 0.65, BLEU-2: 0.47, BLEU-3: 0.33, BLEU-4: 0.23

based model where they suggested a custom CNN architecture responsible for extracting image features and utilising Bidirectional Long Short Term Memory (BiLSTM) as a decoder for caption generation. They have trained and evaluated their model using their own dataset called Bangla natural language image to text (BNLIT). They have achieved a BLEU-1 score of 0.65 after evaluating their model on the test dataset.

Khan et al. [5] have suggested an end-to-end image captioning system where ResNet-50 is for image feature extraction and one dimensional CNN for generating captions, and they used BanglaLekhaImageCaption dataset to train and test their model. Their proposed system achieved a BLEU-1 score of 0.65.

Humaira et al.[3] have presented a performance evaluation of Bangla captioning systems using pre-trained models such as InceptionResnetV2, Xception as encoders and BiLSTM or BiGRU as decoders while using the BanglaLekhaImageCaptions dataset. They have achieved a maximum BLEU-1 score of 0.674 after evaluating their model.

Palash et al. [7] have provided a novel transformer-based architecture that automatically generates Bangla captions from an input image. They have proposed a new transformer architecture with an attention mechanism as a decoder and employed ResNet-101 as an encoder. They have trained and evaluated the model using the BanglaLekhaImageCaptions dataset and achieved a BLEU-1 score of 0.69.

From the previous related works, it is evident that there is a clear scope for further improvement in BLEU score of the Bangla image captioning model. In addition, we did not find any attempt to deploy the model on an embedded device.

In this research work, an encoder-decoder network for Bangla image caption generation and explore different pre-trained image feature extractors for the encoder network is proposed. Finally, we deploy the Bangla image captioning model with the best BLEU score onto an embedded device.

III. DATASET

In the research work, the BanglaLekhaImageCaptions dataset proposed by Rahman et al. [12] is utilized. The downloadable dataset is available online in Mendeley Data[3]. This dataset includes photos with Bengali annotations.

All of its captions are annotated by native Bengali people. There are only two captions tagged with each image in this dataset, yielding a total of 18308 descriptions for the 9154 images. BanglaLekha has 5270 distinct Bengali words. All popular picture captioning datasets are primarily influenced by western culture, with the majority of annotations performed in English. Using such datasets to train an image captioning system for Bangla is not effective. Thus, requiring the necessity for a culturally significant dataset in Bengali to generate acceptable image captions from images related to Bangladeshi and greater sub-continental culture. From the dataset, 80% data is used to train the model, and after training, the remaining 20% is used to evaluate and validate the model.

[3]https://data.mendeley.com/datasets/rxxch9vw59/2

IV. PROPOSED SYSTEM

In this section, you present the proposed model architecture for Bangla image captioning.

We have designed an encoder-decoder based model architecture. Fig. 1 shows a high level overview of the model architecture. As we are working with both image data as input and text data as output, extraction features from image and application of word embedding to the text data is required.

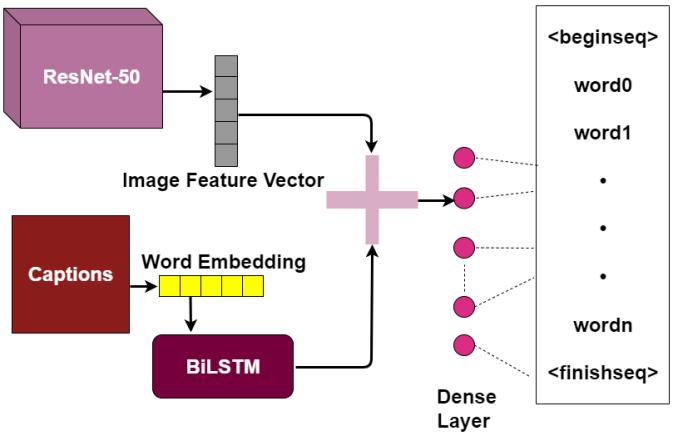


Fig. 1. Overview of Caption Generation Process.

Feature Extraction

A key component of image captioning is the extraction of visual features. We are deploying a pre-trained model called ResNet-50, which has already been trained on millions of images from the ImageNet dataset [19]. As the model is solely utilised for feature extraction, the final two layers have been eliminated, leaving the GlobalAveragePooling layer instead of

a dense layer as the final layer. In contrast to the MaxPooling layer, which generates a 2D matrix, the GlobalAveragePooling layer generates a vector with a dimestion of (None, 2048). The input shape for ResNet-50 is (224,224,3). Therefore, all the photos are reshaped to match this dimension. In the input shape, 3 specifies the number of channels, since the images are in RGB format, the channel number is set to 3.

Word Embedding

Before passing words to RNNs like LSTMs or BiLSTMs, they must be embedded, which turns words into vectors. The embedding layer makes it possible to turn each word into a vector of fixed length and size. The generated vector is dense and contains real values as opposed to merely 0s and 1s. The fixed size of word vectors is the key reason for expressing words with fewer dimensions and in a more efficient manner. In this manner, the embedding layer functions as a lookup table, where the words are the keys and the word vectors are the values. This embedding task is accomplished using the embedding layer of the Tensorflow framework. Using an embedding layer, rather than manually setting values for each word, the embedding values are learned during training. The input and output shapes of this layer are (None, 39) and (None, 39, 128), respectively.

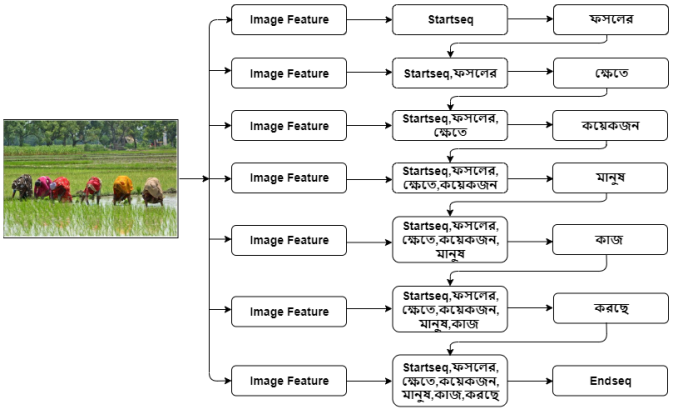


Fig. 2. Example of Word Sequence Generation.

Generation of Word Sequence

Each image in the BanglaLekhaImageCaptions dataset contains only two captions. The maximum and minimum word lengths are 39 and 26, respectively. Although a decrease in word count tends to result in a higher evaluation score [3], the goal of this research work is to develop meaningful, descriptive captions for real-life scenarios, so we use 39 as the fixed word length. During training, zero-padding is employed to increase the length of sentences that are shorter than the fixed maximum length. In addition, a beginseq token and a finishseq token are appended to each pair sequence for identification purposes throughout the training phase. In the training phase, the picture features are extracted from images and the next word in the series is generated using word vectors. Fig. 2 depicts the input-output pair.

In consideration of the limitations of RNNs, LSTMs are a superior option for word generation [20]. However, LSTMs only learn from prior words; for creating syntactically and grammatically accurate sentences, it is also necessary to preserve the knowledge of succeeding words. Therefore, the suggested model uses BiLSTMs, which retains the knowledge learned in both directions, i.e., from both preceding and succeeding words. Fig. 3 depicts the data-flow in BiLSTMs, where P_0,...,P_n are the input words and Q_0,...,Q_n are the outputs of the BiLSTMs which are determined by Eq. 1, where Q_i is output at i^{th} time when activation function h is utilized to weight W_Q and bias B_Q taking into account for forward activation m_i and backward activation n_i at i^{th} time.

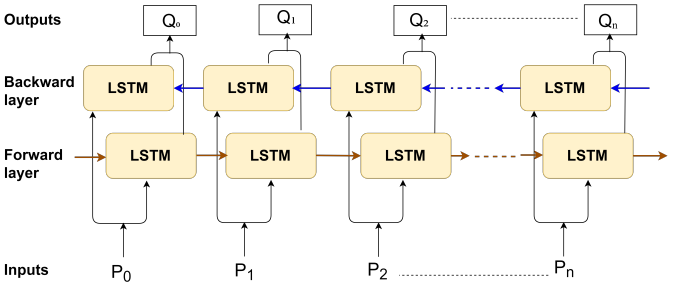


Fig. 3. Illustration of BiLSTMs having P_0,...,P_n as Inputs and Q_0,...,Q_n as Outputs.

$$Q_i = h(W_Q[m_i, n_i] + B_Q) \tag{1}$$

Each LSTM in the BiLSTMs comprises of three gates: input, forget, and output gate. The input gate indicates what incoming information will be stored in the cell state. The forget gate determines what information to discard from the cell state, whereas the output gate provides output at i^{th} time. The corresponding equations for these gates are Eq. 2, Eq. 3, Eq. 4, respectively.

$$j_i = \sigma(W_j[H_{i-1}, P_i] + B_j) \tag{2}$$

$$k_i = \sigma(W_k[H_{i-1}, P_i] + B_k) \tag{3}$$

$$l_i = \sigma(W_l[H_{i-1}, P_i] + B_l) \tag{4}$$

Here, j_i, k_i, l_i is the input, forget and output gate, sigmoid function is represented by σ, W_j, W_k, W_l are the corresponding gate's weights, H_{i-1} is considered to be previous LSTMs block's output at time $i-1$, P_i is the input at i^{th} time and B_j, B_k, B_l are the corresponding gate's bias.

Encoder

The encoder consists of two components, one for managing image feature vectors and the other for managing word sequences. ResNet-50 [21] is used to extract image features originally. These image features are transferred first to a dense layer with 128 units and then to a RepeatVector layer. The RepeatVector layer repeats the inputs for a predetermined number of times. The input to this layer is (None, 128). The RepeatVector's output shape, however, is (39, 128), as

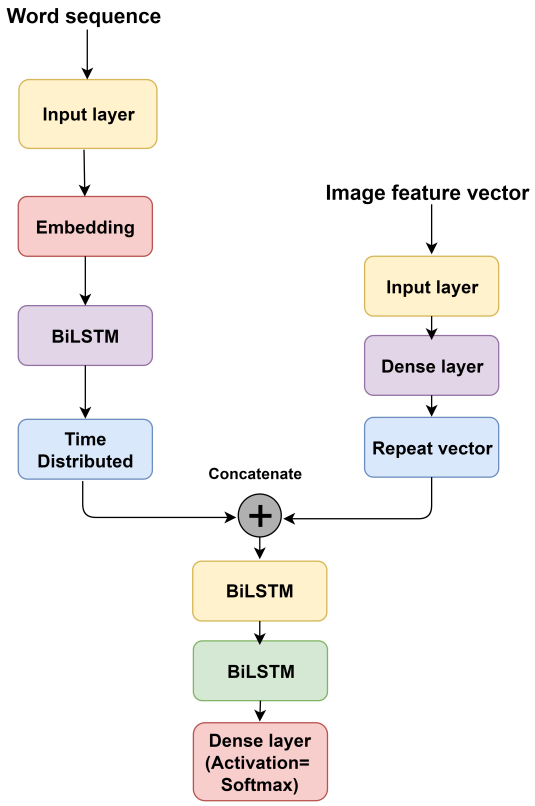


Fig. 4. Encoder-Decoder Model Architecture.

the inputs need to be repeated 39 times to match the output dimension of the other half of the encoder, which handles word vectors. The encoder's word vector processing side includes an embedding layer, a BiLSTM layer with 512 units, and finally a TimeDistributed layer. In contrast with the regular LSTMs, input travels in both directions, and knowledge from each side can be utilised in BiLSTMs. Additionally, it is a potent instrument for modelling the sequential relationships between words and phrases in each direction. In short, BiLSTMs add an additional LSTMs layer that reverses the flow of information, which indicates that the input sequence streams in reverse in the second LSTMs layer. The TimeDistributed layer uses a specified layer (a dense layer with 128 units in the suggested model) for each input vector. Both sides of the encoder have the same output shape, which is (39,128). Their outputs are concatenated and sent to the decoder.

Decoder

The decoder comprises of two BiLSTMs layers and a dense layer. The decoder sends the combined output of the encoder to the first BiLSTMs with 256 units, and the output of the first BiLSTMs is fed to the second BiLSTMs with 512 units. The output of the second BiLSTMs is finally sent to a dense layer with a softmax activation function for word prediction. The model architecture of the encoder-decoder is shown in Fig. 4.

V. Experiments

In this section, a detailed discussion of the experimental details during training and validation of the proposed model architecture for Bangla image captioning is provided.

Experimental Setup

During training of the model, the hardware configuration comprised of a Ryzen 7 3700x CPU, 16GB DDR4 RAM, and Nvidia GTX 1070 8GB graphics card. We implemented our encoder-decoder model using the Tensorflow 2.6 deep learning framework within the Python 3.8 programming language. We deployed the trained Bangla image captioning model onto a Raspberry Pi 4 model B with 8GB RAM.

Parameter Setting

During training, $categorical\_crossentropy$ is used as loss function. Batch size is set to 485. Moreover, $RMSprop$ is selected as the optimizer. $RMSprop$ optimizer selects a distinct learning rate for each parameter during training, which significantly increases model performance. The weights of the model are updated following the Eq. 5 and Eq. 6 during training while using $RMSprop$ optimizer.

$$v_t = \beta_{t-1} + (1-\beta) * {g_t}^2 \tag{5}$$

$$W_{new} = W_{old} - \frac{n}{\sqrt{v_t + \epsilon}} * g_t \tag{6}$$

Here v_t is the average movement speed of gradient, g_t is the cost, β is the moving parameter and in the proposed model its value was 0.99. To calculate the new weights W_{new}, we subtract learning rate (η) times cost g_t which is divided by root over sum of v_t and a constant ϵ of very small value, from the old weights W_{old}.

Performance Metrics

To evaluate the performance of the proposed Bangla image captioning model, we calculate the BLEU score for the generated caption by the trained model. An individualised N-gram BLEU score is the assessment of matching grammes of a particular order, whereas cumulative BLEU scores relate to the computation of single n-gram scores for all orders from 1 to n. Consequently, the cumulative BLEU score is the most reliable metric for evaluating the real-world performance of a sentence generation algorithm. A cumulative BLEU score greater than 70 indicates that the sentence generated by the machine resembles a caption provided by a human. Calculation of cumulative BLEU score is given in Eq. 7, where c refers to length of predicted sentence and r refers to length of the original sentence and p stands for precision.

$$BLEU = min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{4} \frac{logP_n}{4} \tag{7}$$

VI. Result Analysis

In this section, the results found during the training and validation of tge proposed encoder-decoder model for Bangla image is presented. Moreover, we summarise the findings of our research work.

We trained our encoder-decoder model for 90 epochs. Fig. 5 and Fig. 6 exhibits the accuracy curve and the loss curve,

respectively, and it is apparent from both curves that the model converges after 70 epochs. After 70 epochs of training, the proposed model attained an accuracy of 90% and a loss of less than 0.2.

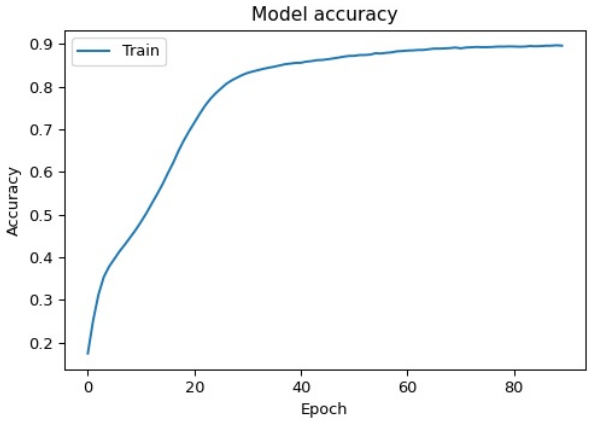


Fig. 5. Accuracy vs Epoch Curve.

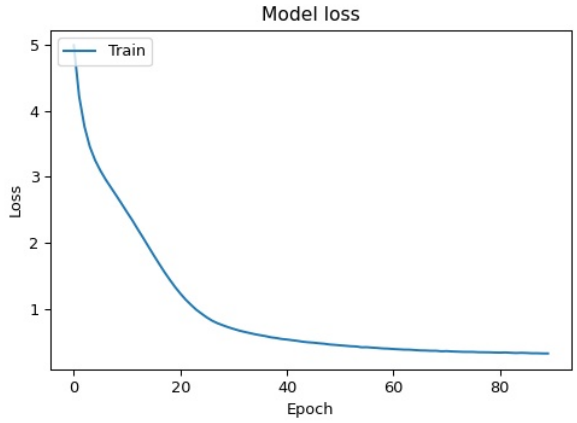


Fig. 6. Loss vs Epoch Curve.

We have implemented three popular pre-trained feature extractors namely VGG-16, Xception, and ResNet-50. Table II shows the BLEU score on the test dataset for the trained models using different pretrained feature extractors. It is found that ResNet-50 performs best as a feature extractor as compared to the VGG-16 or Xception model. The BLEU scores shown in Table II are cumulative BLEU scores.

From Fig. 7, it is evident that our encoder-decoder model with ResNet-50 pretrained model provides a significant improvement in BLEU score compared to the BLEU score stated in the previous research work on the same BanglaLekhaImageCaptions dataset. All of the machine-generated captions displayed in Table III are generated using test samples from the BanglaLekhaImageCaption dataset, which are unseen to the model during the training phase.

Table IV demonstrates that ResNet-50 outperforms the other two feature extractors when evaluated on new data from real-world scenarios. When VGG-16 and Xception are used

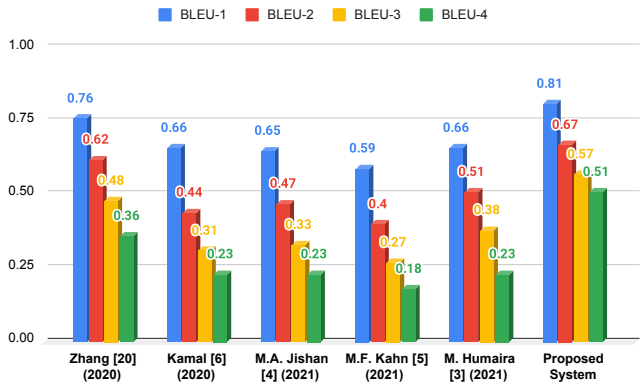


Fig. 7. Accuracy Comparison of Proposed System with Other Research.

as feature extractors, it is observed that the generated captions lack meaning and do not correspond to the meaning of the input image.

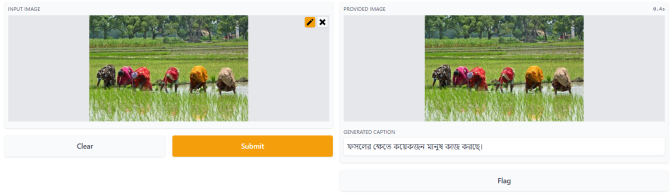


Fig. 8. Web Integration.

For demonstration purposes, we have implemented the model in a web application. This model can also be used for information translation, to assist blind individuals in comprehending their surroundings by converting generated text to speech, and for many other purposes. Fig. 8 depicts an interface for Bangla image captioning within a primary web portal. When a web application receives an image as input, the image is sent to the back-end for image processing and text production. The resulting caption is then shown in the output textbox.

Finally, we have deployed our model on a Raspberry Pi 4 model B with 8GB of RAM. From Table V it is observed that the model only requires 122Mb of storage and the average inference time after testing 100 images was 400ms for ResNet-50 as feature extractor. In addition, the system consumes only 1000mA of current at full load. On the contrary, although flash occupation and energy consumption are similar for all three models, when ResNet-50 is used as a feature extractor, the model takes the least amount of average inference time to generate a caption.

VII. Conclusion

In this research, we have successfully developed and deployed an encoder-decoder based deep learning model named "CapNet" that can generate syntactically and semantically correct and relevant Bangla captions from an input image and deployed the model into an embedded device. The model is trained on a public dataset named BanglaLekhaImageCaption

TABLE II. CUMULATIVE BLEU SCORES COMPARISON ON BANGLALEKHAIMAGECAPTIONS DATASET OF VARIOUS PRE-TRAINED MODEL USED IN ENCODERS OF THE PROPOSED MODEL

Experimental Models	BLEU-1 score	BLEU-2 score	BLEU-3 score	BLEU-4 score
VGG-16 with encoder-decoder	0.45	0.38	0.34	0.31
Xception with encoder-decoder	0.38	0.32	0.29	0.27
ResNet-50 with encoder-decoder	**0.81**	**0.67**	**0.58**	**0.51**

TABLE III. CAPTIONS GENERATED USING PROPOSED MODEL ON TEST-SET OF BANGLALEKHAIMAGECAPTION DATASET

Input Image	Generated Caption
	একজন পুরুষ রিক্সা চালিয়ে যাচ্ছে। (A man is driving a rickshaw.)
	কয়েকজন ছেলে ও একজন মেয়ে নৌকায় বসে আছে। (A few boys and a girl are sitting in the boat.)
	রাস্তায় অনেকগুলো রিক্সা চলছে যেগুলোতে অনেকগুলো মানুষ উঠে আছে। (There are many rickshaws on the road with many people on them.)
	কয়েকটি শিশু পানিতে লাফ দিচ্ছে। (A few children are jumping into the water.)

TABLE IV. CAPTIONS GENERATION COMPARISON USING THREE PRE-TRAINED MODELS AS FEATURE EXTRACTOR IN THE ENCODER

Input Image	ResNet50 incorporated model	Xception incorporated model	VGG-16 incorporated model
	তিনজন বন্ধু মিলে একসাথে দাঁড়িয়ে আছে। (Three friends are standing together.)	একজন পুরুষ সারা শরীরে দিয়েছে। (A man gave the whole body.)	২ জন ছেলে একটি লাইব্রেরীতে দাড়িয়ে ছবি তুলছে (2 boys are standing in a library taking pictures)
	অনেকগুলো নারী ও পুরুষ একসাথে আছে। (Many men and women are together.)	একটি মূর্তি আছে। (There is a statue.)	কিছু ছেলেমেয়ে একসাথে দাড়িয়ে এবং বসে আছেন একটি স্কুলের জায়গায়। (Some children are standing and sitting together in a school place.)
	কিছু মানুষ একটি ক্লাস এর ছাদে দাড়িয়ে এবং বসে আছেন ছবি তোলার জন্য। (Some people are standing and sitting on the roof of a classroom to take pictures.)	ঘরের সামনে কয়েকজন নারী ও কয়েকজন পুরুষ দাড়িয়ে আছে। (A few women and a few men are standing in front of the house.)	অনেকগুলো পুরুষ গোল হয়ে বসে আছে। (Many men are sitting in a circle.)
	পাশাপাশি তিনজন পুরুষ দাঁড়িয়ে আছে। (Three men stand side by side.)	একজন পুরুষ বসে ছবি তুলছে। কয়েকজন বালিকা দাঁড়িয়ে আছে (A man is sitting and taking pictures. Some girls are standing)	একজন বয়স্ক পুরুষ ও একজন নারী আছে। (There is an old man and a woman.)
	একজন নারী বসে আছে। (A woman is sitting.)	দুইজন পুরুষ একজন পুরুষকে এ পানিতে নেমে আছে (Two men are pushing a man into the water)	২ জন ছেলে এবং ১ জন হাতে একটি ব্যাগ নিয়ে বসে আছে একটি ফুলের বেঞ্চে একটি কপালে বই ধরে চেয়ে আছে। (2 boys and 1 with a bag in hand are sitting on a flower bench holding a book on their forehead.)

which has 9154 images with two captions per image. The model has attained the highest cumulative BLEU scores compared to all the previous works on Bangla image captioning that utilised this dataset. In addition, a comparison among three pre-trained image feature extraction models namely ResNet-50, VGG-16, and Xception is provided for the encoder network and it is found that ResNet-50 yields the best BLEU score for Bangla image captioning. Finally, we have deployed our model on a Raspberry Pi 4 model B with 8GB RAM and analysed the inference time and power consumption. In the future, we will enlarge the Bangla image captioning dataset by collecting and labelling more image data and training our model for achieving a higher BLEU score. We will deploy the trained Bangla image captioning model into an Android application so that visually impaired people can have a better understanding of their surroundings.

TABLE V. INFERENCE TIME, ENERGY CONSUMPTION AND FLASH OCCUPATION OF THE PROPOSED MODEL

Model	Flash occupancy (in Mb)	Inference time of proposed system (in ms)	Energy consumption (in mJ)
ResNet-50 with encoder-decoder	122	400	1000
VGG-16 with encoder-decoder	121	1100	1000
Xception with encoder-decoder	119	800	1000

ACKNOWLEDGMENT

We appreciate the financial support provided by the Institute of Energy, Environment, Research, and Development (IEERD, UAP) and the University of Asia Pacific.

REFERENCES

[1] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 2011, pp. 220–228.

[2] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.

[3] M. Humaira, S. Paul, M. Jim, A. S. Ami, and F. M. Shah, "A hybridized deep learning method for bengali image captioning," *IJACSA*, vol. 12, no. 2, pp. 698–707, 2021.

[4] M. A. Jishan, K. R. Mahmud, A. K. Al Azad, M. R. Ahmmad, B. P.

Rashid, and M. S. Alam, "Bangla language textual image description by hybrid neural network model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 2, pp. 757–767, 2021.

[5] M. F. Khan, S. Sadiq-Ur-Rahman, and M. S. Islam, "Improved bengali image captioning via deep convolutional neural network based encoder-decoder model," in *Proceedings of International Joint Conference on Advances in Computational Intelligence*. Springer, 2021, pp. 217–229.

[6] A. H. Kamal, M. A. Jishan, and N. Mansoor, "Textmage: The automated bangla caption generator based on deep learning," in *2020 International Conference on Decision Aid Sciences and Application (DASA)*. IEEE, 2020, pp. 822–826.

[7] M. A. H. Palash, M. Nasim, S. Saha, F. Afrin, R. Mallik, and S. Samiappan, "Bangla image caption generation through cnn-transformer based encoder-decoder network," *arXiv preprint arXiv:2110.12442*, 2021.

[8] Y. Ming, N. Hu, C. Fan, F. Feng, J. Zhou, and H. Yu, "Visuals to text: A comprehensive review on automatic image captioning," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 8, pp. 1339–1365, 2022.

[9] G. Luo, L. Cheng, C. Jing, C. Zhao, and G. Song, "A thorough review of models, evaluation metrics, and datasets on image captioning," *IET Image Processing*, vol. 16, no. 2, pp. 311–332, 2022.

[10] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *Advances in neural information processing systems*, vol. 24, 2011.

[11] R. Lebret, P. Pinheiro, and R. Collobert, "Phrase-based image captioning," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2085–2094.

[12] M. Rahman, N. Mohammed, N. Mansoor, and S. Momen, "Chittron: An automatic bangla image captioning system," *Procedia Computer Science*, vol. 154, pp. 636–642, 2019.

[13] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[14] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1549–1557.

[15] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "Stair captions: Constructing a large-scale japanese image caption dataset," *arXiv preprint arXiv:1705.00823*, 2017.

[16] V. Jindal, "Generating image captions in arabic using root-word based recurrent neural networks and deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[17] S. R. Laskar, R. P. Singh, P. Pakray, and S. Bandyopadhyay, "English to hindi multi-modal neural machine translation and hindi image captioning," in *Proceedings of the 6th Workshop on Asian Translation*, 2019, pp. 62–67.

[18] A. Jaffe, "Generating image descriptions using multilingual data," in *Proceedings of the second conference on machine translation*, 2017, pp. 458–464.

[19] 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*. Ieee, 2009, pp. 248–255.

[20] T. Deb, M. Z. A. Ali, S. Bhowmik, A. Firoze, S. S. Ahmed, M. A. Tahmeed, N. Rahman, and R. M. Rahman, "Oboyob: A sequential-semantic bengali image captioning engine," *Journal of Intelligent & Fuzzy Systems*, vol. 37, no. 6, pp. 7427–7439, 2019.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

An Improved K-Nearest Neighbor Algorithm for Pattern Classification

Zinnia Sultana[1], Ashifatul Ferdousi[2], Farzana Tasnim[3] and Lutfun Nahar[4]
Dept. of Computer Science & Engineering
International Islamic University Chittagong
Chittagong, Bangladesh[1,2,3,4]

***Abstract*—This paper proposed a "Locally Adaptive K-Nearest Neighbor (LAKNN) algorithm" for pattern exploration problem to enhance the obscenity of dimensionality. To compute neighborhood local linear discriminant analysis is an effective metric which determines the local decision boundaries from centroid information. KNN is a novel approach which uses in many classifications problem of data mining and machine learning. KNN uses class conditional probabilities for unfamiliar pattern. For limited training data in high dimensional feature space this hypothesis is unacceptable due to disfigurement of high dimensionality. To normalize the feature value of dissimilar metrics, Standard Euclidean Distance is used in KNN which s misguide to find a proper subset of nearest points of the pattern to be predicted. To overcome the effect of high dimensionality LANN uses a new variant of Standard Euclidian Distance Metric. A flexible metric is estimated for computing neighborhoods based on Chi-squared distance analysis. Chi-squared metric is used to ascertains most significant features in finding k-closet points of the training patterns. This paper also shows that LANN outperformed other four different models of KNN and other machine-learning algorithm in both training and accuracy.**

***Keywords*—LANN algorithm; Standard Euclidian Distance; variance based Euclidian Distance; feature extraction; pattern classification**

I. Introduction

Nearest neighbor classifier is a simplest, oldest and wide-ranging method for classification. It classifies an unidentified pattern by choosing the adjacent example in the training set and measured by a distance metric. It is one of the most common instance-based learning method. Simplicity, transparency and fast training time are the advantage of this algorithm. Instances of nearest neighbor denoted as a point of Euclidian space. It is a conceptual method that can be used to approximate real-valued or discrete-valued target function. K nearest neighbor algorithm is best suited for small data sets and which datasets have less features. This algorithm considers close relationship for similar things. In other words, the similar things of neighbors are considered one of them. For example, if mangoes' appearances is more similar to apple, orange, and guava (fruits) than horse, dog and cat (animals), then most likely mango is a fruit.

In pattern recognition problem, a feature vector $x = (x_1, ——X_q) \epsilon \Re_q$, is considered as an object like J classes, and the goal is to form a classifier that allots x to the exact class from a given set of N training samples. The simplest and alluring approach to solve this problem is the K Nearest Neighbor (KNN) [1][2] classification. Rather than fixed data points this method works on continuous and overlapping neighborhoods [3]. This method uses different neighborhood for each single query so that all points in the neighborhood are adjacent to the query to the extent possible [4][5][6]. KNN uses Straight Euclidean distance to discover the k-closest points from query point [7][8][9][10]. This can influence a real less important feature more than that of others to classify a pattern and misclassify the pattern due to dissimilar metric in measuring the feature values [11][12]. It can seriously affect in the training set with high dimensional feature space [13]. Several biases are introduced in KNN for high dimensional input feature space with limited samples [14].

A modified metric of Standard Euclidean Distance is proposed here, which uses the variance of each feature to give identical influence on the decision to all dissimilar metrics in the feature values [15]. Distance is weighted as chi-squared metric that discovers most relevant features in finding k-closet points to the pattern under consideration from the training space [16].

A locally adaptive form of nearest neighbor classification (LANN) is proposed here to upgrade the obscenity of dimensionality [17]. An effective metric is used here to compute neighborhoods which determines the local decision boundaries from centroid information, and then shrink neighborhoods in directions orthogonal to these local decision boundaries, and extend them parallel to the boundaries [18][19] [20].

To give all features equal influences on the pattern classification a variance based Euclidean distance metric is used in the proposed algorithm instead of straight Euclidean distance metric. The variance of each feature is calculated during training.

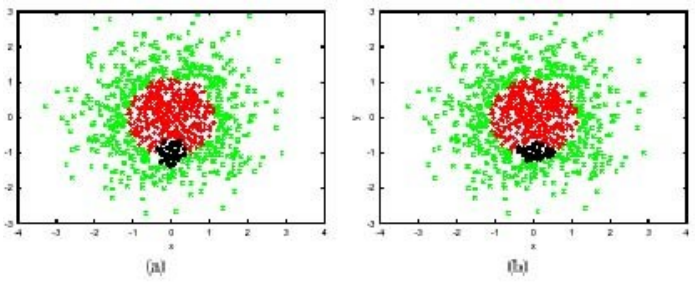


Fig. 1. Neighborhood of the Query Point.

Fig. 1 shows an example. There are two classes and both classes data are produced from a bivariate standard normal

distribution. The radius of class one data is less than or equal to 1.15, while radius of class two is greater than 1.15. As a result, class one is surrounded by class two. Fig. 1(a) shows the nearest neighborhood of size 50 of a query located at (0, -1) near the class boundary. This neighborhood is computed using the Euclidian distance metric Fig. 1(b) displays the neighborhood of same size computed by using the adaptive nearest neighbor classification algorithm. The amended neighborhood is elongated along the direction of the true decision boundary and constricted along the direction orthogonal to it, which is the most relevant direction for the given query.

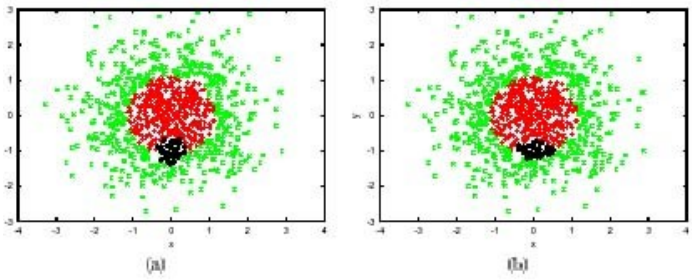


Fig. 2. Spherical Neighborhood of the Query Point.

Fig. 2 Plot (a) shows the spherical neighborhood of the query point (0, -1) containing 50 points (shown as darker circles). Plot (b) shows the corresponding neighborhood found by the proposed algorithm also containing 50 points. After applying the adaptive procedure, the neighborhood is constricted along the most relevant dimension and elongated along the less important one.

This paper proposed an algorithm that can be used in many practical applications of pattern recognition problem in machine learning technology for pattern classification tasks. It has been compared experimentally with KNN, DANN and C4.5 in a large number of artificial and natural learning domains. Experimental result shows that use of Variance based Euclidean distance metric and FRW perfectly removes the problem of constant class conditional probabilities in KNN and improves the performance of KNN.

II. Literature Review

Locally adaptive KNN algorithms indicate the value of k that should be used to categorize an interrogation by accessing the outcomes of cross-validation calculations in the resident locality of the query [21] [22]. Local KNN procedures are exposed to complete analogous to KNN in experimentations with twelve frequently secondhand data sets.

Deepti et al. [23] proposed a Quad Division prototype for stirring uneven class distribution by using Selection based K-nearest Neighbor classifier. Here the performance of QDPS based on KNN technique is assessed in fraud detection in mobile advertising. The utility of the QDPSKNN is likened with base model KNN and other selection methods, namely NearMiss-1, NearMiss-2, NearMiss-3, and Condensed Nearest Neighbor (CNN).

Suyanto et al. [24]introduced a new variant of KNN called Multi-Voter Multi-Commission Nearest Neighbor to observe the profit by enhancing the Local Mean based Pseudo Nearest Neighbor. MVMCNN is gained extra nearby than LMPNN. And then compared it with two single voter models: KNN and BMFKNN, however it shows the multi voter model better decision than the other model.

Armand et al. [25] proposed a metaheuristic search algorithm named Simulated Annealing, to choose an optimal k, thus rejecting the prospect of an exhaustive search for optimal k. Hence, the result is compared with in four different classification method to determine a substantial development in the computational competence compared to the KNN methods.

D. Maruthi et al. [26] introduced an effective classification system for MRI brain tumor and for giving grade of brain tumor images. The images are classified by using the adaptive k nearest neighbor classifier. However, the classification and segmentation arrival method are valued by accuracy, sensitivity and specificity.

Jieying et al. [27] proposed a precise image interpolation with adaptive KNN for searching image on the input image patch and conduct them for nonlinear mapping among low resolution and high-resolution image patches.

Jianping et al. [28] offered a local mean representation based k nearest neighbor classifier to increase the performance of classification and exceed the primary issues of KNN classification. They used two databases UCI and KEEL and also three common databases that carried out by liken LMRKNN and KNN based. However, it shows the LMRKNN significantly outperforms the KNN based methods.

Some previous works on K-Nearest Neighbor Algorithm for Pattern Classification that we have discussed above (Table I). Apart from this, no such similar topic related work exists as far as our knowledge. Our primary focus is to propose an algorithm that can be used in many practical applications of pattern recognition problem in machine learning technology for pattern classification tasks. It has been compared experimentally with KNN, DANN and C4.5 in a large number of artificial and natural learning domains but there is no work found that use the comparison among AI and NLP domain. Besides, no relation is shown in any research as per our study with use of Variance based Euclidean distance metric and FRW which perfectly removes the problem of constant class conditional probabilities in KNN and improves the performance of KNN.

III. Methodology

LANN has three main components: Variance-based Euclidean distance Metric, Feature Relevance Weight (FRW), the best K value using the majority voting scheme [12] [13]. LANN uses a variance based Euclidean Distance metric to find the adjacent neighbors of a query point from the training space and then the class is assigned with the majority class of the neighbors. The component of each feature in the distance is normalized using the variance. While finding the nearest points, distance component of each feature is weighted with chi-squared distance metric to work out the most relevant features.

The main steps of the algorithm and the working procedure are as follows (Fig. 3):

TABLE I. RELATED WORKS ON LAKNN

Ref. No	Description	Model	Limitation
[23]	Proposed a Quad Division Prototype Selection based K-nearest Neighbor classifier for establishing stirring uneven class distribution.	QDPSKNN, PS method	This method is not works well over real time large sized datasets.
[24]	Introduced a new variant of KNN called MVMCNN which is planned to observe the profits by enhancing LMPNN.	KNN, MVMCNN, LMPNN, BMFKNN	It does not give complete inquiries for the definite datasets.
[25]	To choose an optimal K value proposed a metaheuristic search algorithm and also eliminate the prospect of an exhaustive search	KNN, Adaptive algorithms, Parameter Optimization	The adaptive KNN method can't achieve good performance.
[26]	Introduced an effective classification system to classify MRI brain tumor	AKNN, Median filter	Can't provide an explanation in optimization computation complexity problem.
[27]	Proposed an accurate image interpolation with adaptive KNN searching and nonlinear regression.	AKNN	Do not explore deep learning models.
[28]	Proposed a k nearest neighbor classifier based on local mean representation.	KNN, LMRKNN	Can't explore deep learning models.
[29]	Introduced a method named density based adaptive k nearest neighbor.	Nearest Neighbor Classification, Density based method DBANN	can't create extra artificial examples to recompense for smaller class
[30]	An adaptive procedure monitoring system was planned base2d on the KNN rule.	KNN	This method does not suitable for simple processes.

Step-1: Start several Leave-One-Out Tests (Test index "T") for a single neighbor (T=1) to a threshold value (T=10). For each Leave-One-Out test, each example in the training space is classified according to the step 2 to 7.

Step-2: For each test point x0 in training space in each leave-one-out test (Query point index "j" of each "T" value, Given input parameters: K0, K1, K2, L), Initialize a feature relevance weight "wi" to 1 for each feature component in Euclidian distance measure in equation 1.

$$D(x,y) = \sqrt{\sum_{i=i}^{q} w_i \frac{x_i - y_i^2}{variance(i^{th} feature}} \tag{1}$$

$$Variance(ithfeature) = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N} \tag{2}$$

$\overline{x}$is the mean value of ith feature, where q is the number of features of each point. x, y are the two data points and distance between x and y data point is D (x, y). xi and yi are the ith and ith feature value of x and y data point respectively. Equation 1 measures Euclidean distance with the normalized weight of each feature according to the variance of that feature that are in training data set. wi is the feature relevance weight for each feature.

Step-3: Compute K0 nearest neighbors of x0 by means of the variance based weighted Euclidian distance metric using equation 1 for $w_i = 1$.

Step 4: For each feature i, i =1…..q, compute feature relevance measure through equation 3 to equation 7.

$$\overline{r_i}(x_0) = \frac{1}{k} \sum_{z \varepsilon N(x_0)} r_i(z) \tag{3}$$

where $N(x_0)$ represents the neighborhood of x_0 holding the K_0-nearest training point.$r_i(x_0)$ denotes the capability of feature i to predict $Pr(j|z)s$ at $x_i = z_i$ and defined as follows:

$$r_i(z) = \sum_{j=1}^{J} \frac{[Pr(j|z) - \overline{Pr}(j|x_i = z_i)]^2}{\overline{Pr}(j|x_i = z_i)} \tag{4}$$

The nearer $\overline{Pr}(j|x_i = z_i)$ is to $Pr(j|z)$, the additional information features i carries for predicting the class posterior probabilities locally at z.$\overline{Pr}(j|x_i = z)$ is the conditional expectation of $p(j|x)$, given that xi assumes value z, where xi represents the ith feature of x.$Pr(j|z)$ and $\overline{Pr}(j|x_i = z_i)$ is estimated as follows:

$$Pr(j|z) = \frac{\sum_N^{n=1} 1(x_n \varepsilon N_i(z)) 1(y_n = j)}{\sum_N^{n=1} 1(x_n \varepsilon N_1(z))} \tag{5}$$

1(.) is function which acts as indicator, such that if the argument is true it returns 1 and if false then returns 0. $N_1(z)$ is the neighborhood centered at z containing K_1 nearest training points.

$$\overline{Pr}(j|x_i = z_i) = \frac{\sum_{x_n \varepsilon N_2(z)} 1(\mid x_{ni} - z_i \mid \leq \triangle_i) 1(y_n = j}{\sum_{x_n \varepsilon N_2(z)} 1(\mid x_{ni} - z_i \mid \leq \triangle_i)} \tag{6}$$

$N_2(z)$ is the neighborhood centered at z containing K_2 nearest training points, the value of $\triangle_i$ is selected from the interval containing a fixed number of L points:

$$\sum_{n=1}^{N} 1(\mid x_{ni} - z_i \mid \leq \triangle_i) 1(x_n \epsilon N_2(z)) = L \tag{7}$$

Step 5: Update Feature Relevance Weight (FRW) "w_i" according to equation 8 to equation 9. Feature Relevance Weight (FRW) is calculated by:

$$w_i(x_0) = (R_i(x_0))^t / \sum_{q}^{l=1} (R_i(x_0))^t \tag{8}$$

where $R_i(x_0)$ is defined by

$$R_i(x_0) = (max) Tj(r_i)(x_0) - (r_i)(x_0) \tag{9}$$

t = 1, 2 giving quadratic weighting scheme. In all our experiments we obtained optimal value for input parameters $K_1 = 5$, $K_2 = 10\%$ of N, K_0= 15% of N. L is set to half of the K_2.

Step 6: Iterate steps 2 to 5 again, in this situation each feature has some FRW value.

Step 7: Using Step 2 to 6 a FRW for each feature is obtained. Using FRW in variance based Euclidian distance metric; distance of all examples in training data set with query point x0 is calculated. The examples are ordered in ascending according to their distance value. Among them, a total of "T" examples are chosen from lowest distance to T th point. The class value with maximum number of examples is taken as the class value (majority voting scheme) of the query point x_0.

Step 8: All examples in the training space are classified following the steps from 2 to 7.

Step 9: Error rate is calculated for Tth Leave-One-Out test.

Step 10: All (T=10) Leave-One-Out Tests are completed and error rate is recorded for each test. Test with minimum error rate is chosen as best k-value for the training data set.

Step 11: Using the best k-value; classify any query point following the steps from 2 to 7.

The algorithm of LANN appears to be complex, but the core of LANN is the application of three main components: Variance based Euclidean distance metric, Feature relevance weight, Choice of the best k-value.

Algorithm (Dtraining, x0)

INPUT: $D_{training}$: a set of training examples.

x_0: a query point to be predicted.

OUTPUT: A predicted class value for x_0.

q=No. of Features in the training data set.

N=Total no. of Example in Training dataset.

for T=1 to threshold value (T=10) **do**
  for j=1 to N **do**
    x_j=An example from $D_{training}$
    D= $D_{training}$ - x_j
    Initialize FRW w_i=1 //Label-1//
    for m=1 to 2 **do**
      P=compute weighted distance of x_j by the equation 1 from D.
      $N(x_j)$ =Sort the examples(D) in ascending on P and choose K_0 neighbors from lowest distance.
      Q=compute weighted distance of $z \varepsilon N(x_j)$ by the equation 1 from D.
      $N_1(z)$ = Sort the examples(D) in ascending on Q and choose K_1 neighbors from lowest distance.
      $N_2(z)$ = Sort the examples(D) in ascending on Q and choose K_2 neighbors from lowest distance.
      for each dimension i , i=1...q, compute Relevance Measure $\overline{r_i}(x_i)$ through equation 3 to 7. **do**
        Update FRW w_i according to the equation 7 to 9.
      end for
    end for
    Compute weighted distance of x_j using the new FRW w_i by equation 1 from D. //Label-2 //
    E=Choose "T" neighbors from D Apply majority voting on E and classify x_j.
  end for
  Calculate error rate for "T" test.
end for
K = Choose best T with lowest error rate.
Compute a FRW for x_0 following steps from "Label-1 to Lablel-2".
F = Choose "K" neighbors from $D_{training}-x_0$ Class C =Apply majority voting on "F"
Return "C"

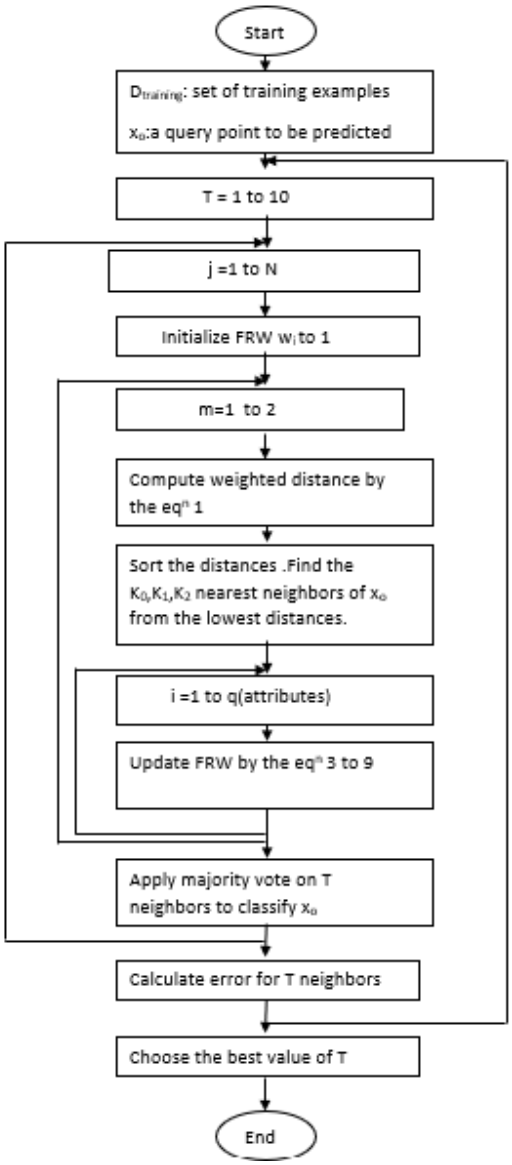


Fig. 3. Flowchart of LANN.

IV. Experimental Result

Twelve different real data sets are studied for experimental analysis of LANN. The Breast Cancer, Iris, Diabetes, Glass,

Vowel, Sonar, Hepatitis, Wine, Segmentation, Lymphography, Liver-Disorder and Lung-Cancer data are taken from UCI Machine Learning Repository [4]. All for the datasets we perform Leave-One-Out test to measure performance (Table II).

TABLE II. DOMAINS USED IN THE PROPOSED ALGORITHM (LANN)

Description of the domains used in experimental study.			
Domain name	Size	No. of classes	No. of Attributes
Breast Cancer	699	2	9
Iris	150	3	4
Diabetes	768	2	8
Glass	214	6	9
Vowel	528	11	10
Sonar	208	2	60
Hepatitis	150	2	19
Wine	178	3	13
Segment	2310	7	19
Lymphography	148	4	18
Liver-Disorder	345	3	6
Lung-Cancer	32	3	56

Table III shows the leave one out test result for 12 datasets. Table III depicts the Leave-One-Out error rates for the four methods under consideration on the twelve real world data.

The above table shows error rates (%) for different K-values. Column 1 of Table III shows that the minimum error rate is 2.43 for K=4 in breast cancer dataset. Column 2 of Table III shows minimum error rate 3.33 for K=2 for Diabetes dataset, minimum error rate for Iris dataset is 3.33 that shown in column 3 of Table III, So, the best K-value is 6. Minimum error rate for Glass dataset is 24.76 for k value 4 is shown in column 4 of Table III, 9.13 is the minimum error rate for k value 4 for sonar dataset shown in column5 of Table III, for k value 2 minimum error rate 0.56 is found for Vowel dataset that shown in column 6 of Table III, column 7 of Table III shows the minimum error rate of Hepatitis dataset which is 21.33 for k value 2. Minimum error rate of Wine dataset, Segment dataset, Lymphographic dataset, Liver disorder dataset and Lung Cancer dataset is 1.68 for k value 2, 1.63 for k value 4, 8.10 for k value 2, 22.31 for k value 4, 37.5 for k value 4 are shown in column 8, 9,10,11 and 12 respectively of Table III.

After completion of all Leave-One-Out tests we calculate the error rate of LANN by the following:

$$Errorrate(\%) = \frac{No.of\backslash failures" * 100}{TotalNo.ofInstances}$$

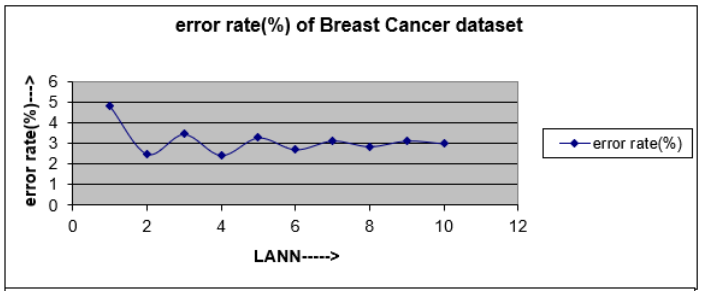


Fig. 4. Error Rate (%) Graph for Breast-Cancer Dataset.

"Fig. 4"' shows the error rate (%) graph for Breast-Cancer dataset for C4.5, DANN, KNN, LANN where the error rates

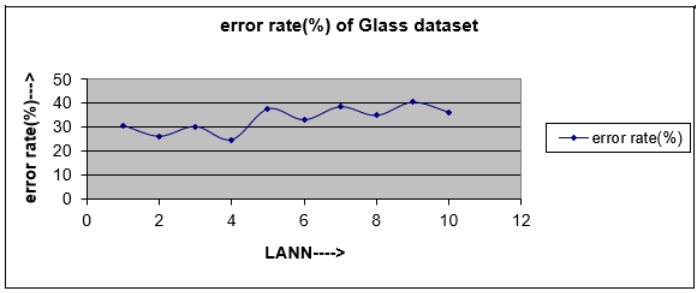


Fig. 5. Error Rate (%) Graph for Glass Dataset.

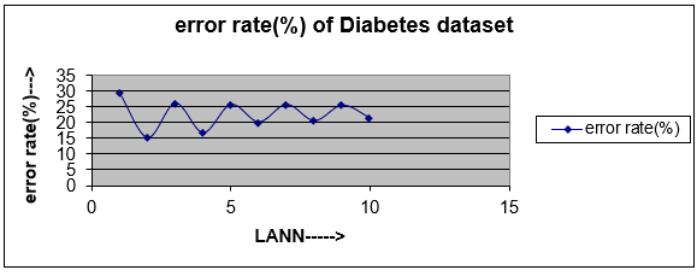


Fig. 6. Error Rate (%) Graph for Diabetes Dataset.

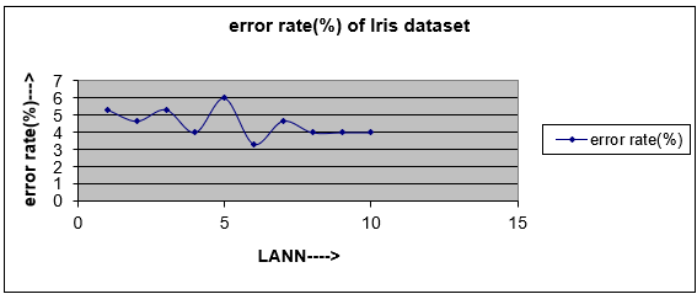


Fig. 7. Error Rate (%) Graph for Iris Dataset.

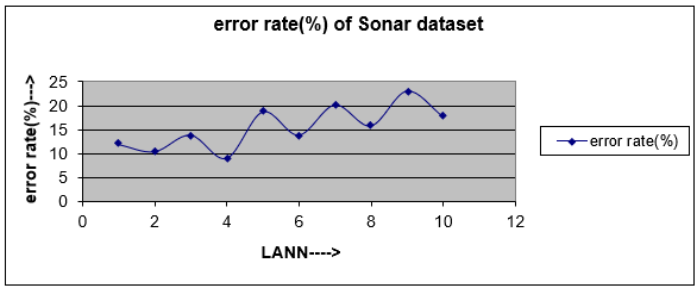


Fig. 8. Error Rate (%) Graph for Sonar Dataset.

4.70, 3.10, 4.10, 2.43 respectively. "Fig. 5" is for Diabetes dataset where the error rate is 25.00 for C4.5, 18.10 is for DANN, 24.40 is for KNN and 15.10 is for.Error rates for Iris dataset is shown in "Fig. 6". For this dataset the error rate is 8.00 for C4.5, 6.00 for DANN, 8.00 for KNN and for LANN it is 3.33; Glass dataset's error rate is shown in "Fig. 7" where the error rates for C4.5, DANN, KNN, LANN are 31.80, 27.10, 28.00, 24.76 respectively.Error rates for Sonar dataset is shown in "Fig. 8" which shows 23.10, 7.70, 12.50, 9.13 for

TABLE III. THE LEAVE-ONE-OUT TEST RESULTS FOR 12 DATASETS ARE GIVEN BELOW

	Breast cancer	Diabetis	Iris dataset	Glass dataset	Sonar dataset	Vowel dataset	Hepatitis dataset	Wine dataset	Segment dataset	Lympho graphy dataset	Liver Disor-der dataset	Lung Cancer Dataset
1	4.86	29.16	5.33	30.84	12.01	0.75	40.00	3.37	3.10	21.62	37.68	50.00
2	2.80	15.10	4.66	25.54	10.50	0.56	21.33	2.80	1.90	8.10	23.23	38.10
3	3.86	26.04	5.33	30.37	13.94	2.08	33.33	3.93	2.82	14.18	35.65	65.62
4	2.43	16.66	4.00	24.76	9.13	1.51	26.66	1.68	1.63	9.45	22.31	37.50
5	3.29	25.65	6.00	37.85	18.75	5.68	32.00	2.80	3.19	17.56	35.94	65.62
6	2.71	20.05	3.33	33.17	13.94	3.97	26.00	1.68	2.90	13.51	24.63	46.87
7	3.14	25.65	4.67	38.78	20.19	8.71	30.00	3.37	4.42	18.91	39.71	70.00
8	2.86	20.44	4.00	35.04	15.86	7.00	26.00	2.81	3.10	15.54	26.95	68.75
9	3.14	25.39	4.00	40.65	23.07	13.06	32.86	2.81	3.15	18.91	37.39	72.50
10	3.00	21.35	4.00	36.44	17.78	10.22	30.00	2.24	3.12	16.89	29.85	70.50

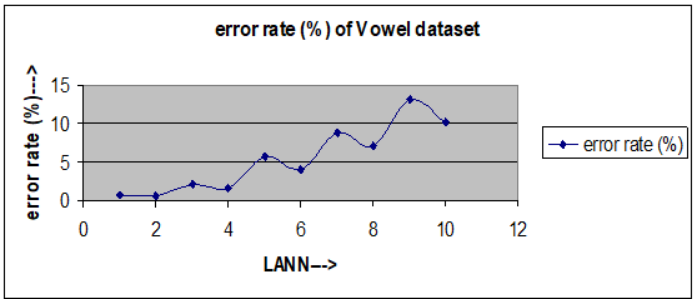


Fig. 9. Error Rate (%) Graph for Vowel Dataset.

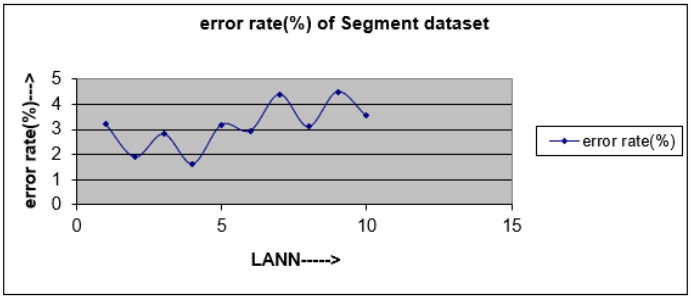


Fig. 12. Error Rate (%) Graph for Segment Dataset.

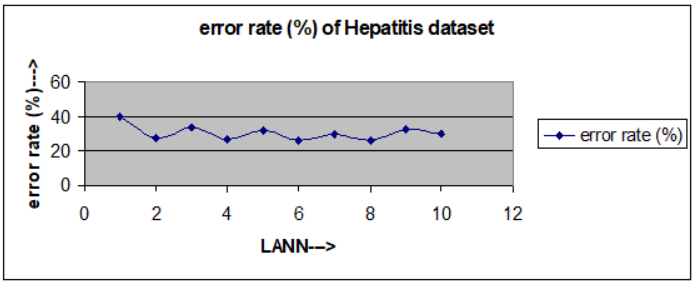


Fig. 10. Error Rate (%) Graph for Hepatitis Dataset.

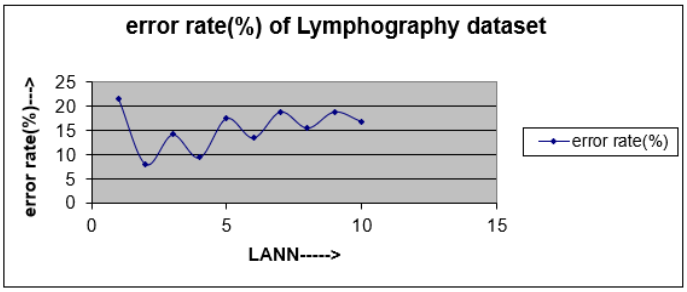


Fig. 13. Error Rate (%) Graph for Lymphography Dataset.

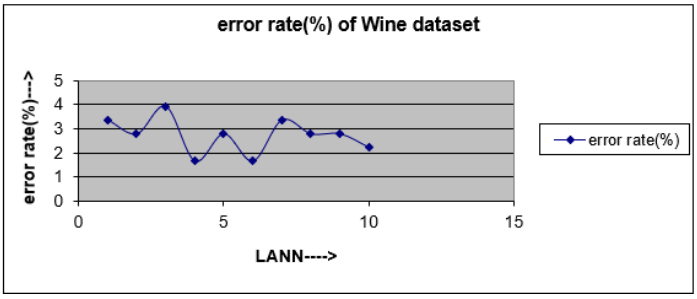


Fig. 11. Error Rate (%) Graph for Wine Dataset.

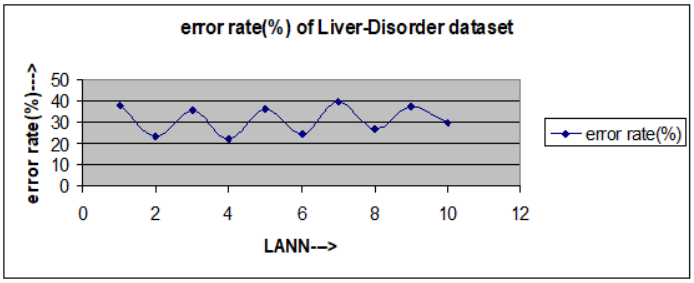


Fig. 14. Error Rate (%) Graph for Liver-Disorder Dataset.

four algorithms. “Fig. 9” shows the error rate for Vowel dataset for C4.5, DANN, KNN, LANN where error rates are 36.70, 12.50, 11.80, 0.56 respectively. Error rate (%) for Hepatitis dataset is shown in “Fig. 10” where the error rate is 18.40 for C4.5, 20.40 is for DANN, 22.30 is for KNN and 21.33 is for LANN. Wine datasets error rate is shown in “Fig. 11” the error rate (%) for C4.5 is 12.10, for DANN error rate is 13.50,14.60 is for KNN and 1.68 is for LANN. “Fig. 12” the error rate (%) where the error rates are 3.70, 2.50, 3.60, 1.63 for C4.5, DANN, KNN, LANN respectively. From “Fig. 13” error rates has been observed for Lymphography dataset where the error rate is 21.90 is is for C4.5, 17.70 for DANN, 19.30 is for KNN and 8.10 is for proposed LANN. Error rates for Liver-Disorder dataset is shown in “Fig. 14” where the error

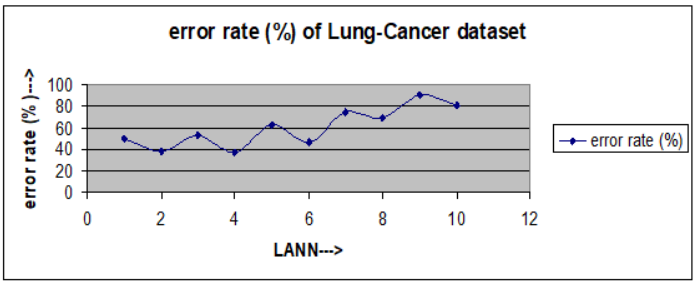


Fig. 15. Error Rate (%) Graph for Lung-Cancer Dataset.

rates are 35.10, 32.30, 34.50, 22.31 for C4.5, DANN, KNN, LANN respectively. "Fig. 15" depicts thr error rate for Lung-cancer dataset, where the error rate for for C4.5 is 57.50, for DANN it is 45.90, for KNN it is 47.90 and for LANN it is 37.50.

From the comparison Table IV it is observed that the average error rate (%) of proposed algorithm (LANN) is 12.32 whereas the average error rate (%) for C4.5, DANN, KNN are 23.17, 17.23, 19.10. Thus it can be said that, the efficiency of LANN is better than the above algorithms (Fig. 16).

TABLE IV. COMPARISON OF LANN W. R. TO OTHER ALGORITHMS

Domain no.	Domain name	C4.5	DANN	KNN	LANN
1	Breast cancer	4.70	3.10	4.10	2.43
2	Diabetes	25.00	18.10	24.40	15.10
3	Iris	8.00	6.00	6.00	3.33
4	Glass	31.80	27.10	28.00	24.76
5	Sonar	23.10	7.70	12.50	9.13
6	Vowel	36.70	12.50	11.80	0.56
7	Hepatitis	18.40	20.40	22.30	21.33
8	Wine	12.10	13.50	14.60	1.68
9	Segment	3.70	2.50	3.60	1.63
10	Lymphography	21.90	17.70	19.30	8.10
11	Liver Disorder	35.10	32.30	34.50	22.31
12	Lung-Cancer	57.50	45.90	47.90	37.50
	Average	23.17	17.23	19.10	12.32

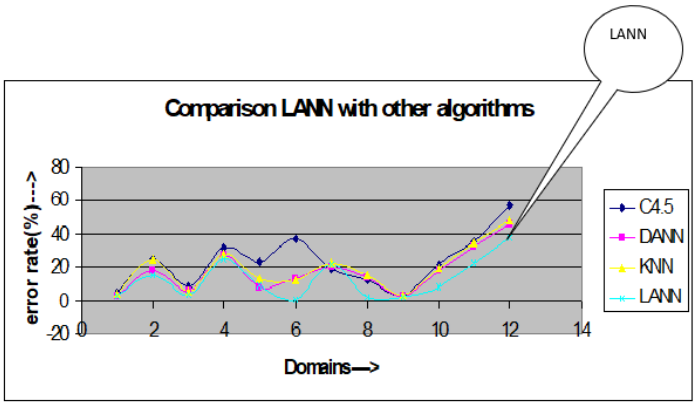


Fig. 16. Error Rate (%) of Different Domains. Horizontal Axis gives the Domain's no. and the Vertical Axis gives the Corresponding Error Rate.

V. DISCUSSIONS

There are basically two parts for pattern classification. By using an algorithm the first part creates feature vector from a given image and these features are used in the second part to learn a machine to classify an unknown pattern.

These two parts are not completely independent, this means machine learning algorithms may be benefited by knowing how the features are extracted from an image and feature extraction may be more fruitful if the type of machine leaning algorithm is known. However, the limitation of this paper is it only explored second part. That is, this work emphasis on to build a system which can classify an unknown image or pattern by using machine learning from a given set of database, all of which feature vectors have already been broken down into by an image processing algorithm. For example, the Segment dataset that is used in this work is an image classification problem. After applying the proposed algorithm (LANN) on the Segment dataset, the classification error rate is observed as 1.6%, whereas the error rates for C4.5, DANN, KNN are 3.7%, 2.5%, 3.6%, respectively. It proves that the LANN performs better than other existing algorithms in image-classification problems (Fig. 17.)

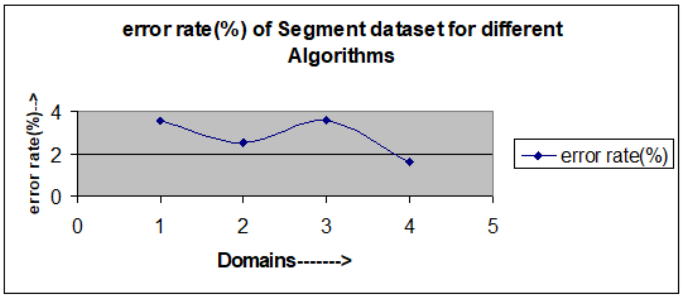


Fig. 17. Comparison of Error Rates (%) of Proposed Algorithm (LANN) with Other Algorithms for Segment Dataset.

VI. CONCLUSION

LANN presents a new variant of nearest neighbor method to classify pattern effectively. To produce neighborhood, it uses a flexible metric that are elongated along less relevant feature dimensions and constricted along most influential ones. By using this technique, the class conditional probabilities tend to be more homogeneous in the modified neighborhoods. From the experimental result it is clearly shown that LANN can potentially improve the performance of K-NN and recursive partitioning methods in some classification problems. The results are also in favor of LANN over other adaptive methods such as C4.5 and DANN.

REFERENCES

[1] A. J. Gallego, J. R. Rico-Juan, and J. J. Valero-Mas, "Efficient k-nearest neighbor search based on clustering and adaptive k values," *Pattern Recognition*, vol. 122, p. 108356, 2022.

[2] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of k-nearest neighbour (knn) algorithm and its different variants for disease prediction," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022.

[3] R. R. Rajammal, S. Mirjalili, G. Ekambaram, and N. Palanisamy, "Binary grey wolf optimizer with mutation and adaptive k-nearest neighbour for feature selection in parkinson's disease diagnosis," *Knowledge-Based Systems*, vol. 246, p. 108701, 2022.

[4] "Journal of Ambient Intelligence and Humanized Computing, 12(2), 2867-2880. UCI Repository of Machine Learning Databases: ," http://www.ics.uci.edu/ mlearn/MLRepository.html, accessed: 2010-09-30.

[5] B.-W. Yuan, X.-G. Luo, Z.-L. Zhang, Y. Yu, H.-W. Huo, T. Johannes, and X.-D. Zou, "A novel density-based adaptive k nearest neighbor method for dealing with overlapping problem in imbalanced datasets," *Neural Computing and Applications*, vol. 33, no. 9, pp. 4457–4481, 2021.

[6] W. Zhu, W. Sun, and J. Romagnoli, "Adaptive k-nearest-neighbor method for process monitoring," *Industrial & Engineering Chemistry Research*, vol. 57, no. 7, pp. 2574–2586, 2018.

[7] S. S. Mullick, S. Datta, and S. Das, "Adaptive learning-based k-nearest neighbor classifiers with resilience to class imbalance," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5713–5725, 2018.

[8] B. Tu, S. Huang, L. Fang, G. Zhang, J. Wang, and B. Zheng, "Hyperspectral image classification via weighted joint nearest neighbor and sparse representation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 4063–4075, 2018.

[9] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[10] J. H. Friedman *et al.*, "Flexible metric nearest neighbor classification," Citeseer, Tech. Rep., 1994.

[11] I. Gazalba, N. G. I. Reza *et al.*, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. IEEE, 2017, pp. 294–298.

[12] C. Patgiri and A. Ganguly, "Adaptive thresholding technique based classification of red blood cell and sickle cell using naíve bayes classifier and k-nearest neighbor classifier," *Biomedical Signal Processing and Control*, vol. 68, p. 102745, 2021.

[13] L. Xiong and Y. Yao, "Study on an adaptive thermal comfort model with k-nearest-neighbors (knn) algorithm," *Building and Environment*, vol. 202, p. 108026, 2021.

[14] J. Zheng, W. Song, Y. Wu, and F. Liu, "Image interpolation with adaptive k-nearest neighbours search and random non-linear regression," *IET Image Processing*, vol. 14, no. 8, pp. 1539–1548, 2020.

[15] D. Jiang, W. Zang, R. Sun, Z. Wang, and X. Liu, "Adaptive density peaks clustering based on k-nearest neighbor and gini coefficient," *IEEE Access*, vol. 8, pp. 113 900–113 917, 2020.

[16] A. V. Kachavimath, S. V. Nazare, and S. S. Akki, "Distributed denial of service attack detection using naïve bayes and k-nearest neighbor for network forensics," in *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)*. IEEE, 2020, pp. 711–717.

[17] Z.-X. Guo and P.-L. Shui, "Anomaly based sea-surface small target detection using k-nearest neighbor classification," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 6, pp. 4947–4964, 2020.

[18] H. Su, Y. Yu, Z. Wu, and Q. Du, "Random subspace-based k-nearest class collaborative representation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 6840–6853, 2020.

[19] B. Wang and Z. Mao, "A dynamic ensemble outlier detection model based on an adaptive k-nearest neighbor rule," *Information Fusion*, vol. 63, pp. 30–40, 2020.

[20] J. Gou, W. Qiu, Z. Yi, Y. Xu, Q. Mao, and Y. Zhan, "A local mean representation-based k-nearest neighbor classifier," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 3, pp. 1–25, 2019.

[21] H. Ye, P. Wu, T. Zhu, Z. Xiao, X. Zhang, L. Zheng, R. Zheng, Y. Sun, W. Zhou, Q. Fu *et al.*, "Diagnosing coronavirus disease 2019 (covid-19): Efficient harris hawks-inspired fuzzy k-nearest neighbor prediction methods," *Ieee Access*, vol. 9, pp. 17 787–17 802, 2021.

[22] A. Lin, Q. Wu, A. A. Heidari, Y. Xu, H. Chen, W. Geng, C. Li *et al.*, "Predicting intentions of students for master programs using a chaos-induced sine cosine-based fuzzy k-nearest neighbor classifier," *Ieee Access*, vol. 7, pp. 67 235–67 248, 2019.

[23] D. Sisodia and D. S. Sisodia, "Quad division prototype selection-based k-nearest neighbor classifier for click fraud detection from highly skewed user click dataset," *Engineering Science and Technology, an International Journal*, vol. 28, p. 101011, 2022.

[24] S. Suyanto, P. E. Yunanto, T. Wahyuningrum, and S. Khomsah, "A multi-voter multi-commission nearest neighbor classifier," *Journal of King Saud University-Computer and Information Sciences*, 2022.

[25] J. Zhang, T. Wang, W. W. Ng, and W. Pedrycz, "Knnens: A k-nearest neighbor ensemble-based method for incremental learning under data stream with emerging new classes," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[26] Y. Wang, X. Cao, and Y. Li, "Unsupervised outlier detection for mixed-valued dataset based on the adaptive k-nearest neighbor global network," *IEEE Access*, vol. 10, pp. 32 093–32 103, 2022.

[27] Y. Cai, J. Z. Huang, and J. Yin, "A new method to build the adaptive k-nearest neighbors similarity graph matrix for spectral clustering," *Neurocomputing*, vol. 493, pp. 191–203, 2022.

[28] A. Onyezewe, A. F. Kana, F. B. Abdullahi, and A. O. Abdulsalami, "An enhanced adaptive k-nearest neighbor classifier using simulated annealing," *International Journal of Intelligent Systems and Applications*, vol. 13, pp. 34–44, 2021.

[29] D. M. Kumar, D. Satyanarayana, and M. Prasad, "Mri brain tumor detection using optimal possibilistic fuzzy c-means clustering algorithm and adaptive k-nearest neighbor classifier," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 2867–2880, 2021.

[30] Z. Pan, Y. Pan, Y. Wang, and W. Wang, "A new globally adaptive k-nearest neighbor classifier based on local mean optimization," *Soft Computing*, vol. 25, no. 3, pp. 2417–2431, 2021.

An Approach to Detect Phishing Websites with Features Selection Method and Ensemble Learning

Mahmuda Khatun[1]
Department of CSE
Comilla University
Cumilla - 3506, Bangladesh

MD Akib Ikbal Mozumder[2]
Department of CSE
Comilla University
Cumilla - 3506, Bangladesh

Md. Nazmul Hasan Polash[3]
Department of CSE
Comilla University
Cumilla - 3506, Bangladesh

Md. Rakib Hasan[4]
Dept. of Information & Comm. Technology
Comilla University
Cumilla - 3506, Bangladesh

Khalil Ahammad[5]
Department of CSE
Comilla University
Cumilla - 3506, Bangladesh

MD. Shibly Shaiham[6]
Department of CSE
CUET
Chattagram-4349, Bangladesh

***Abstract*—Nowadays, phishing is a major problem on a global scale. Everyone must use the internet in today's society in order to cope up in the real world. As a result, internet crime like phishing has become a serious issue throughout the world. This type of crime can be committed by anyone; all they need is a computer. Additionally, hacking may now be learned quickly by anyone with programming and mathematical skills. The adoption of various techniques by anti-phishing toolbars, such as machine learning, may enable users to quickly identify a fake website. As a result, researchers are now particularly interested in the problem of detecting fraudulent websites. Machine learning techniques have been offered throughout the entire process to more precisely identify fraudulent websites. To find the best accurate outcome, classification with random parameter tuning and ensemble based approaches are utilized. A user-friendly interface has also been suggested to make the system more accessible to the public.**

***Keywords*—Machine learning; deep learning; catboost; LGBM; embedded; react-native; flask**

I. Introduction

Technology has made individuals more dependent than they have ever been before. As the price of electronic devices such as smartphones, tablets, personal computers, laptops, and so on continues to drop, an ever-increasing number of people are able to purchase them and are using them. However, the rise of cyber dangers has been the most significant in recent decades. Phishing, which is responsible individually for 90 percent [1] of the data breaches, is one of the most common methods that people are tricked into giving over their personal information. People who conduct most of their financial transactions and shopping online are the most likely to fall victim to phishing scams. Criminals used to extort money from unsuspecting victims by threatening them with guns, stealing their cars, or by using force. When compared to less developed countries, industrialized nations have a lower incidence of this particular form of criminal activity. But in recent years, phishing has become a significant issue all around the world. This is due to the fact that one does not require any kind of weapon in order to commit this form of crime; all that is required is a computer. In addition, the availability of books and guides to follow on the internet makes it possible for anyone who is mathematically savvy and has some experience with programming to learn how to hack. The Federal Bureau of Investigation estimates that more than 26 billion dollars have been stolen from businesses and individuals around the world. In addition, the number of new websites that are used for phishing is continually expanding. Fig. 1 indicates new phishing websites are increasing year by year which is an immense threat for users. Therefore, cutting-edge research needs to be carried out in order to neutralize this danger. Also many cyber security training are providing by both government or by private companies which rises the awareness of people basically in the form of games [2] to [3].

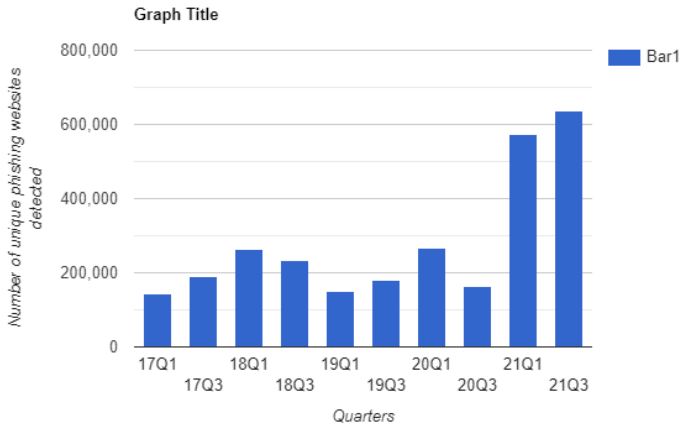


Fig. 1. The Number of New Phishing Websites from First Quarter of 2017 upto Third Quarter of 2021 by Statista.

Internet scams that use email, social media, and websites to steal private information are called phishing. Attackers take advantage of both system flaws and user ignorance. As a result, researchers must design solutions at both the technical and user-level levels. It is possible to protect yourself from Phishing by using a variety of different methods. Anti-phishing toolbars primarily protect users on their computers' local machines from phishing attacks. Games and other awareness campaigns are used to educate people about the issue [4] to[5].

Machine learning, rule-based, visual similarity, and other non-content-based methods are used by anti-phishing toolbars to identify phishing sites. Malicious URL detection technology aids users in spotting malicious links and wards off attacks from harmful websites. So, the first motivation is to give safety to the users. Another motivation is to increase the reliability of such technology. For that, increasing the accuracy of the system is a crucial factor which has been done in this research work. As a result, we conducted a comprehensive study of the current literature on phishing types, anti-phishing techniques, and anti-phishing toolbars in order to better understand the faults of existing toolbars and provide researchers with improved solutions. In order to compile the papers, we used Google Scholar as a database. Supervised Machine Learning with different Features Selections has been used in the past to detect phishing websites [6]. Classifiers that use wrapper-based techniques typically have the most useful features and the best overall performance [7] to [8]. Multi-model simulations are used in this phase for higher performance accuracy [9]. In order to achieve the highest level of accuracy, a hybrid model incorporates the greatest elements of multiple models while minimizing the negatives of each. The most accurate results are obtained using ensemble methods [10] to [11]. Using wrapper feature selection and ensemble learning, a phishing detection software will be created. Wrapping features selection approach delivers the most desired features in order to acquire the most accurate outcome. In addition, an ensemble technique is utilized to find the accuracy of all classifiers and select the best classifier to fit with it in order to forecast phishing websites.

Section II shows the literature review and Section III briefly describes the methodology and design of our study. Section IV elaborates the experimental results and Section V concludes with limitation and some recommendations of future work of the study.

II. Literature Review

Many researchers worked previously on analysing phishing websites. Our technique is to combine their works and get a better result. Also there exists many phishing apps and some are in the form of games which literally use live servers to detect urls [12]. We went through the past works and try to improve the accuracy of models and propose a user friendly interface to detect them.If we want to run a detection app successfully then we have to dig into the process first. M.A Tahir [11] proposed a hybrid model to detect phishing sites using supervised learning algorithm. He used machine learning methods and combine them to find the best accuracy but skip the feature selection methods. Zamir [13] proposed a diverse machine learning method to detect phishing websites wheres they skip the part of using feature extraction methods. Sharivari [14] also proposed a system where they used almost every classification techniques. R.kohavi [7] shows wrapper based features selection classifier work with all the possible features subsets and utilizes a ml classifier as an evaluation function of the features subsets. The highest evaluation is considered as the best feature subset. W.Ali [6] tried to propose a system with machine learning classifier and the feature selection method. He found that random forest with wrapper based feature selection method gives the highest accuracy of 97.3%. The most accurated result found is 97.4%. Our proposed system is to improve their accuracy by using same dataset. Also there are so many apps or games and they mainly use live servers or open source to detect urls [12].Our project is to make a user friendly app with our proposed ml method.

There exists many methods to detect fake urls. Some based on machine learning techniques and skip features selection methods. In our project we use classifiers as well as feature selection methods and also implement this as an application. The researchers applied simple feature identification and extraction techniques. It can also be possible to get more accurate results by using more powerful classifiers in ML such as category boosting and Deep Learning algorithms also can handle this type of problem. The existing models did not extract as much as features like us. So, the reduced number of feature extraction in the existing models is a gap. There is scope to improve the accuracy of the existing models. In the next section we will describe how we overcome this lacking found a better way to overcome some problems.

III. Methodology and Design

Methodology is considered as the best part of any research. It implements the answer of the questions start with “How?”. It’s a tremendous way to ensure the acceptability of the work through valid and reliable results. For this proposed method, we have gone through several steps. At first valid data have been collected. Then this data have gone through several steps till a user interface use this model to predict the urls. Machine learning classifiers combining with feature selection methods have been used and react native is used for given it a real life application.

There are two major steps to illustrate this project. First of all selecting the best features among all features and then select an ensemble machine learning method to get the result.

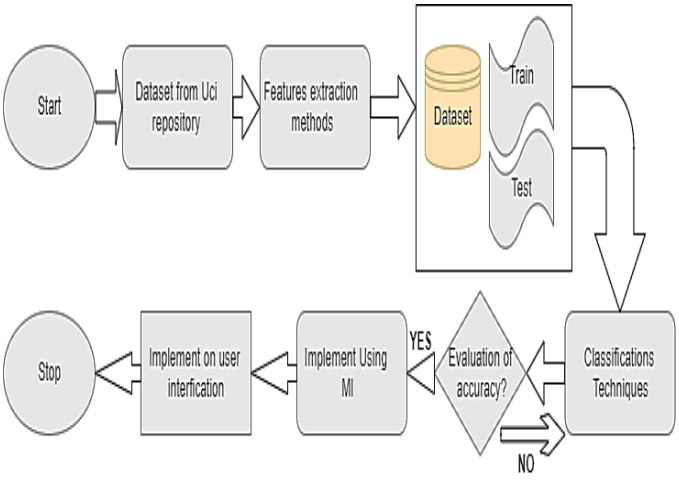


Fig. 2. Flow Chart for whole Process.

Fig. 2 illustrates the methodology of phishing website detection based on supervised machine learning classifiers with features selection method.

Fig. 3 illustrates the classes we have used for Ml model classification and feature extraction methods.They have been rotating to find the best accuracy among them.

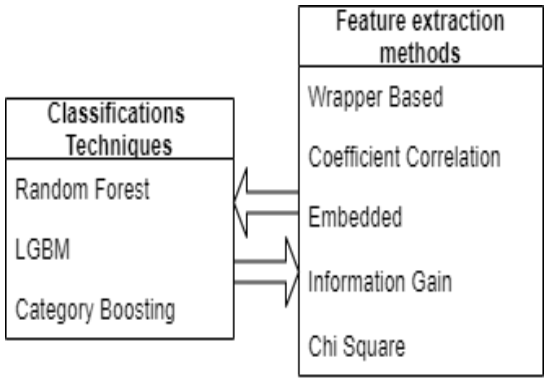


Fig. 3. Loop between Ml Model Classifiers and Features Extraction Technique.

Five steps are required to be accomplished in order to detect the phishing website: dataset collection, features extraction, features selection, training of machine learning classifiers, and evaluation of machine learning classifiers

The dataset of our project were collected from the UCI Machine Learning Repository, which is freely available for use. This dataset consists of 11055 rows and 30 columns which were used to extract several website features [15]

A. Explanation of Proposed Method

The complete explanation of the proposed method is given in this section. The basics of all procedures in the methodologies are discussed.

1) Data Preprocessing and Analysing: For building a good accurate model preprocessing of data is must. Otherwise the model may fail to give correct results. Data processing is a term which is basically processing of raw unusable data to suitable machine data.Using J48 algorithm [16] is very helpful in examine the data categorically and continuously. Fig. 4 shows the preprocessing procedure of the data. Here we have used basic types of data preprocessing methods in our project.

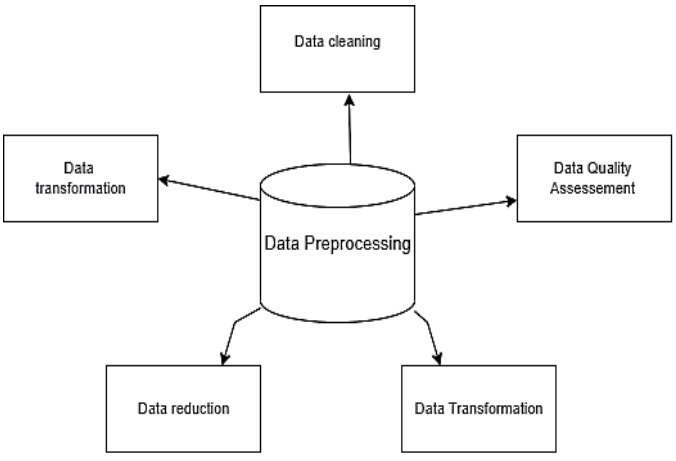


Fig. 4. Data Pre-Processing.

The following are some fundamental data preprocessing methodologies that have been implemented in the proposed method.

1) We have dropped one column which only indicates the index value sequentially for all the records.
2) Then We have analysis the unique classes for each of the data.
3) Next we have replaced all the values of records to 0 or 1
4) Those records who have missing values are also eliminated.

2) Feature Extraction: ML can't work best on choosing all the features. So feature selection method is used for dealing with crude content straightforwardly. It reduces the dimension of feature space.During feature extraction, uncorrelated and useless features will be removed which can help to fit the best useful algorithm and can improve the accuracy of the model

Wrapper Based: Subset of features are used and with each subset, train a model. We add or remove features based on the accuracy given by the subsets.It is also possible to utilize wrapper classification algorithms that combine dimensionality reduction and classification, although their computational costs are considerable and their discriminative power is limited. Furthermore, in order to achieve high accuracy, these strategies rely on the effective selection of classifiers. There are three types of wrapper based selection:

1) Forward Selection
2) Backward Elimination
3) Stepwise Selection

In our proposed method Stepwise selection secured best accuracy among all of the features selection methods.

Correlation Based: High correlated features depends each other and linearly inter dependent, hence have almost the same effect.When a feature expansion doesn't result in an improvement, the algorithm moves on to the next best unexpanded subset. The entire feature subset space is searched by this approach without any restrictions. As a result, there should be a limit on backtracking. The program then returns the feature subset that produced the highest merit up to that point.

Information Gain: The information gained is a system where amount of information improved before splitting them are counted. This is actually the mutual information between two random variables.Determining an attribute's relevancy and, consequently, its position inside the decision-tree, is the main goal of the Information Gain. An attribute (variable) with numerous distinct values prevents the information gain from effectively differentiating among the attributes.

Chi Square: Categories in a dataset are tested using the chi-square method. We determine the Chi-square among each element and the intended outcome and then choose the features with the highest Chi-square scores. It assesses whether the relationship between two categorical data in the sample corresponds to their true relationship in the population.

3) Machine Learning Algorithm: Machine learning is a learning technique and getting new information without explicit instruction with the aid of training data. Basically, there are two different categories of machine learning methods.

1) Supervised

2) Unsupervised

The method of learning from leveled data known as "supervised learning" provides an output for each input. In accordance with this, the algorithms operate. Unsupervised learning refers to the type of learning when the training or operation is carried out on an unlabeled dataset. To discover related features, the algorithm aggregates data of a similar type. Whatever the case, the labeled data is used to develop our model. A supervised learning problem can thus be used to frame the issue. There are two different kinds of supervised learning algorithms.

1) Classification
2) Regression

As we have used Supervised learning technique and dataset is classifed, in the next part we will only discuss about supervised classification methods.

Random Forest: A random forest is a ml method for tackling classification and regression issues. It makes use of ensemble learning, a method for solving complicated issues by combining a number of classifiers.

Numerous decision trees make up a random forest algorithm. The random forest algorithm creates a "forest" that is trained via bagging or bootstrap aggregation. The accuracy of ml algorithms is increased by bagging, an ensemble meta-algorithm.

Based on the predictions made by the decision trees, the (random forest) algorithm determines the outcome. It makes predictions by averaging or averaging out the results from different trees. The accuracy of the result grows as the number of trees increases.

Light Gradient Boosted Machine: A framework and a type of gradient boosting is called a light gradient boosting machine, or LGBM. Light GBM is based on Decision tree methods, just as another gradient boosting method. We can decrease memory utilization and boost efficiency with the aid of Light GBM.

The primary distinction between Light GBM and other gradient boosting frameworks is that Light GBM grows leaf-wise and in a vertical orientation. The other algorithms, however, grow horizontally in a level-wise fashion. The leaf that produces the least inaccuracy and the most efficiency is chosen by Light GBM. This approach is significantly more beneficial in lowering the mistake percentage. In other words, it expands leaf-wise whilst others grow level-wise.

Category Boosting: The CatBoost algorithm plays an essential part for supervised machine learning applications and is based on Gradient Descent. It will work effectively for issues involving categorical data. A series of decision trees are generated sequentially when training this model because the CatBoost technique is based on gradient decision trees. Each new tree that is created as training goes on has a lower loss than the one before it. It is employed for a variety of functions, including weather forecasting, self-driving cars, recommendation systems, personal assistants, and search.

4) React Native APP: Basics of react native: JavaScript provides the framework for React Native, which allows developers to write genuine mobile applications that run natively on iOS and Android devices [17]. It is based on React, which is a JavaScript toolkit that Facebook uses for designing user interfaces; however, rather than targeting browsers, it targets mobile platforms. To put it another way, web developers now have the ability to create mobile applications that have a look and feel that is genuinely "native", and they can do it from the convenience of a JavaScript library that we are already familiar with and adore. In addition, the majority of the code that you create may be shared between platforms, which makes it simple to develop applications for both Android and iOS at the same time using React Native. Fig. 5 indicates that how server and app are interacting with each other.

Architecture of our react native app:

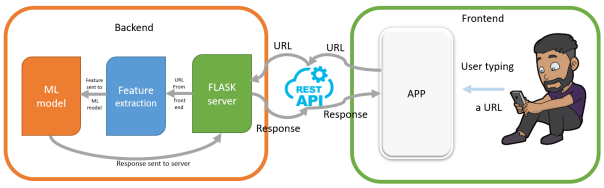


Fig. 5. Architecture of React Native.

The architecture has mainly two parts. Backend and frontend. Each part is divided into some essential parts.

5) Implementation of Machine Learning Model:: The steps required to implement are as follows:

- **Step-1:** At first we collected the data from online [15]
- **Step-2:** Then we did some analysis of the data which we have already discussed.
- **Step-3:** We did some preprocessing of the data that is necessary for building the models. We first balance the dataset as the data were very class imbalanced. Then we classified data and discard the record which have null values
- **Step-4:** Then we did the most important part which is feature extraction. We have used the wrapper based,embedded method, correlation coefficient, information gain and chi square. Then we implemented this techniques one by one with ensemble machine learning techniques.
- **Step-5:** After feature extraction, we have split the dataset into train and test set. We have used 70split from the same dataset.
- **Step-6:** We have built the machine learning model. We have used ensemble models [10] like Random forest, LGBM, Catboosting method. And then We have implemented this methods with features extraction methods which We have discussed already in the previous part.
- **Step-7:** After building the model we evaluated the performance of the model. And then implemented this methods to the user interface.

6) Implementation of User Interface using our Machine Learning Model:: Here we will discuss about the user interface which will be a real life url detector using our machine learning approach [18]. Backened: Here a server is created to communicate with the frontend and ML model. It sends the response of ML model to the frontend.

Flask Server
Flask is a python framework. With the help of Flask, a server was created. It receives URL via REST API from frontend. It is also responsible to send the response of ML model to REST API.

Extracting Features
Features from URL are extracted at this end. The extracted features are then fed into the ML model.

Response form ML Model
With the help of the features, the ML model creates a response whether the URL is phishing or safe.

Frontend: The frontend supplies the URL to the Backend through REST API.

App
The react native app receives input from user. The URL is saved to its state and then it is sent to the REST API. The API sends it to the server.

User
The user gives an URL as input to the App. The user also can see the response of the ML model in the app interface.

IV. Results

Performance are the outcomes and the most important parts of a project. The better a model performs the more useful that model is. In the previous chapters all the steps for the proposed methodology has been explained. Now it's time to evaluate the model. In this section we will describe the overall outcomes and we will show our proposed model with the traditional approaches. The performance and the real life user interface will be discussed here.

A. Dataset Description

The dataset of our project have already discussed in the previous chapter. Dataset were collected from the UCI Machine Learning Repository, which is freely available for use. This dataset consists of 11055 Urls and 30 columns which were used to extract several website features [15].

Fig. 6, 7, 8 and 9 present the symbol count, ip count, url length and shortening url feature, respectively. Their characteristics is put on x axis and on axis respective numbers is kept. Here -1 means the feature has no impact to make it malicious, 0 means neutral and 1 it is used to identify the url as malicious. Here count for some features are given. we have tried this for all of the features which indicate how many individual classes for each of the features.

B. Impact Analysis

Fake url has a great impact on society.It has also a tremendous impact on almost every sectors of life like economy and politics. It has become a curse nowadays. So detecting the fake url is much needed thing to overcome this global issue. Some impacts are shown below:

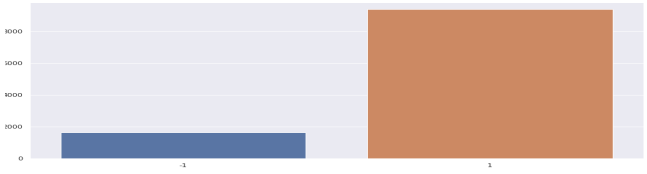


Fig. 6. Symbol Count.

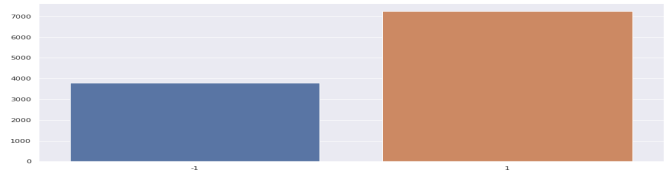


Fig. 7. Ip Count.

1) Social and Environmental Impact: Phishing on social media sites like Instagram, LinkedIn, Facebook, or Twitter is referred to as social media phishing. Such an assault aims to take control of your social media account or to steal personal information. The negative impacts of phishing on a business are numerous and include financial loss, loss of intellectual property, reputational harm, and interruption of daily operations. These outcomes combine to reduce a company's value, sometimes with disastrous results.

2) Ethical Impact: The act of designing and carrying out simulations that are certain to cause stress and anxiety in all levels and roles of your employee base is ethical impact of phishing. The high rate of undesirable actions—such as clicking on a link, opening an attachment, or entering login information based on private or delicate topics—is not being generated on purpose or with predetermined objectives. Without the right context, these strategies are certain to endanger your program. An atmosphere of trust will be difficult to restore, regain, or establish.

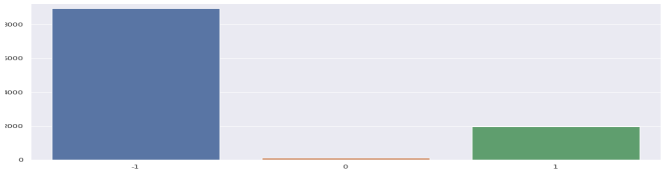


Fig. 8. Url Length.

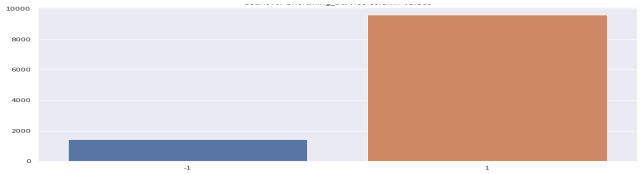


Fig. 9. Shortening Url.

C. Evaluation of Proposed Method

In this section, the method of our propose system will be shown step by step. Here we will discuss about our feature extraction technique, ensemble machine learning and finally the connection with the server.

1) Feature Extraction: For feature extraction procedure we have used wrapper based, embedded method, Correlation and Coefficient, Information Gain, Chi square method. But checking the performance with ml techniques, we have come to a decision to use wrapper based feature extraction. It gives the highest result among all of them [7]. In this dataset, there were thirty columns before using feature extraction method. Fig. 10 shows the columns before extracting features.

```
Index(['having_IPhaving_IP_Address', 'URLURL_Length', 'Shortining_Service',
       'having_At_Symbol', 'double_slash_redirecting', 'Prefix_Suffix',
       'having_Sub_Domain', 'SSLfinal_State', 'Domain_registeration_length',
       'Favicon', 'port', 'HTTPS_token', 'Request_URL', 'URL_of_Anchor',
       'Links_in_tags', 'SFH', 'Submitting_to_email', 'Abnormal_URL',
       'Redirect', 'on_mouseover', 'RightClick', 'popUpWidnow', 'Iframe',
       'age_of_domain', 'DNSRecord', 'web_traffic', 'Page_Rank',
       'Google_Index', 'Links_pointing_to_page', 'Statistical_report'],
      dtype='object')
```

Fig. 10. Before Wrapper Used.

Then using wrapper, we got the performance of each of the subsets and came to know which subsets give the most accurate result which is shown in Fig. 11.

	feature_idx	cv_scores	avg_score	feature_names
1	(7,)	[0.88891904115784 71]	0.888919	(SSLfinal_State,)
2	(7, 13)	[0.91261872455902 3]	0.912619	(SSLfinal_State, URL_of_Anchor)
3	(7, 13, 14)	[0.91895070104025 32]	0.918951	(SSLfinal_State, URL_of_Anchor, Links_in_tags)
4	(5, 7, 13, 14)	[0.92455902306548 57]	0.924559	(Prefix_Suffix, SSLfinal_State, URL_of_Anchor,...
5	(5, 7, 13, 14, 28)	[0.92799638172777 25]	0.927996	(Prefix_Suffix, SSLfinal_State, URL_of_Anchor,...
6	(5, 7, 13, 14, 25, 28)	[0.93423790140208 05]	0.934238	(Prefix_Suffix, SSLfinal_State, URL_of_Anchor,...
7	(5, 6, 7, 13, 14, 25, 28)	[0.94590682948891 9]	0.945907	(Prefix_Suffix, having_Sub_Domain, SSLfinal_St...
8	(5, 6, 7, 13, 14, 23, 25, 28)	[0.95151515151515 15]	0.951515	(Prefix_Suffix, having_Sub_Domain, SSLfinal_St...
9	(1, 5, 6, 7, 13, 14, 23, 25, 28)	[0.95748530076888 28]	0.957485	(URLURL_Length, Prefix_Suffix, having_Sub_Doma...
10	(1, 5, 6, 7, 13, 14, 23, 24, 25, 28)	[0.96182722749886 93]	0.961827	(URLURL_Length, Prefix_Suffix, having_Sub_Doma...
11	(0, 1, 5, 6, 7, 13, 14, 23, 24, 25, 28)	[0.96716417910447 76]	0.967164	(having_IPhaving_IP_Address, URLURL_Length, Pr...
12	(0, 1, 5, 6, 7, 12, 13, 14, 23, 24, 25, 28)	[0.97241067390321 12]	0.972411	(having_IPhaving_IP_Address, URLURL_Length, Pr...
13	(0, 1, 5, 6, 7, 12, 13, 14, 23, 24, 25, 26, 28)	[0.97657168701944 82]	0.976572	(having_IPhaving_IP_Address, URLURL_Length, Pr...
14	(0, 1, 5, 6, 7, 12, 13, 14, 23, 24, 25, 26, 27...	[0.98046132971506 11]	0.980461	(having_IPhaving_IP_Address, URLURL_Length, Pr...
15	(0, 1, 5, 6, 7, 8, 12, 13, 14, 23, 24, 25, 26,...	[0.98317503392130 26]	0.983175	(having_IPhaving_IP_Address, URLURL_Length, Pr...
16	(0, 1, 5, 6, 7, 8, 12, 13, 14, 16, 23, 24, 25,...	[0.98543645409317 05]	0.985436	(having_IPhaving_IP_Address, URLURL_Length, Pr...
17	(0, 1, 5, 6, 7, 8, 12, 13, 14, 15, 16, 23, 24,...	[0.98715513342379 01]	0.987155	(having_IPhaving_IP_Address, URLURL_Length, Pr...
18	(0, 1, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 23,...	[0.98796924468566 26]	0.987969	(having_IPhaving_IP_Address, URLURL_Length, Pr...
19	(0, 1, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 18,...	[0.98869289914066 04]	0.988693	(having_IPhaving_IP_Address, URLURL_Length, Pr...
20	(0, 1, 3, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, ...	[0.98914518317503 4]	0.989145	(having_IPhaving_IP_Address, URLURL_Length, ha...
21	(0, 1, 3, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, ...	[0.98950701040253 27]	0.989507	(having_IPhaving_IP_Address, URLURL_Length, ha...
22	(0, 1, 2, 3, 5, 6, 7, 8, 11, 12, 13, 14, 15, 1...	[0.98959746720940 76]	0.989597	(having_IPhaving_IP_Address, URLURL_Length, Sh...
23	(0, 1, 2, 3, 5, 6, 7, 8, 11, 12, 13, 14, 15, 1...	[0.98968792401628 22]	0.989688	(having_IPhaving_IP_Address, URLURL_Length, Sh...
24	(0, 1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15...	[0.98968792401628 22]	0.989688	(having_IPhaving_IP_Address, URLURL_Length, Sh...
25	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14,...	[0.98968792401628 22]	0.989688	(having_IPhaving_IP_Address, URLURL_Length, Sh...
26	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	[0.98968792401628 22]	0.989688	(having_IPhaving_IP_Address, URLURL_Length, Sh...
27	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	[0.98968792401628 22]	0.989688	(having_IPhaving_IP_Address, URLURL_Length, Sh...
28	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	[0.98968792401628 22]	0.989688	(having_IPhaving_IP_Address, URLURL_Length, Sh...
29	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	[0.98968792401628 22]	0.989688	(having_IPhaving_IP_Address, URLURL_Length, Sh...
30	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	[0.98968792401628 22]	0.989688	(having_IPhaving_IP_Address, URLURL_Length, Sh...

Fig. 11. Subset by Wrapper.

Fig. 12 shows the final outcomes which selected 23 features among all the features discrete rest of them.

2) Machine Learning Model: After using feature selection method, we have used this feature extraction method with several machine learning techniques as like Random forest, Light Gradient Boosted method and Category Boosting. We have come in a decision that Random Forest gives the highest

```
['having_IPhaving_IP_Address',
 'URLURL_Length',
 'Shortining_Service',
 'having_At_Symbol',
 'Prefix_Suffix',
 'having_Sub_Domain',
 'SSLfinal_State',
 'Domain_registeration_length',
 'HTTPS_token',
 'Request_URL',
 'URL_of_Anchor',
 'Links_in_tags',
 'SFH',
 'Submitting_to_email',
 'Redirect',
 'popUpWidnow',
 'age_of_domain',
 'DNSRecord',
 'web_traffic',
 'Page_Rank',
 'Google_Index',
 'Links_pointing_to_page',
 'Statistical_report']
```

Fig. 12. Wrapper Final Selection.

accuracy with Coefficient Correlation. So in the next part, RF model will be discussed.

Fig. 13. Heatmap of RF with Coefficient.

Next in the Fig. 14 shows the highest accuracy done by random forest.

```
[ ]  pred=rf.predict(X_test)
     accuracy_score(y_test, pred)

     0.9746759119686463
```

Fig. 14. Accuracy.

Here in Fig. 13 shows the heatmap of random forest classification. Heatmap actually provides realtime analytics to understand the performance.

Here it shows that RF gives the highest result of 97.47% among all others and best suitable for proposed system.

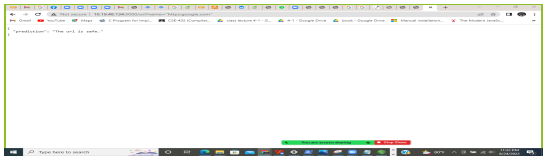


Fig. 15. Connecting with Server.

Finally in Fig. 15 shows our user interface connected with the server system.

D. Evaluation of Performance

In this section we will show a comparative analysis of the performance between our proposed methodologies and

traditional approaches. At first we will show the analysis with different feature selection methods with different machine learning methods done by the base papers. Then we will show the performance of our proposed methods we have tried. Here, at first the performance of the traditional approaches are shown. In Fig. 16 shows the performance done by the previous paper [13]. The highest performance is done by using NN classifier + Random forest + bagging which is about 97.4.

Techniques	Highest achieved accuracy (%)
PWP using RF classifiers (Ibrahim and Hadi (2017)	95.2
Intelligent detection using RF (Subasi *et al.* (2017)	97.3
Proposed method using stacking	97.4
Proposed method using RF	97.3

Note: PWP = phishing web sites prediction

Fig. 16. Performance by Paper [13].

In Fig. 17 shows the performance done by the previous paper [14]. The highest performance is done by using Random forest which is about 97.3.

classifier	train time (s)	test time(s)	accuracy	recall	precision	F1 score
logistic regression	0.080971	0.006414	0.926550	0.943968	0.925700	0.934704
decision tree	0.021452	0.003737	0.965988	0.971414	0.967681	0.969531
random forest	0.436126	0.021941	0.972682	0.981484	0.969852	0.975622
ada booster	0.336519	0.016766	0.936953	0.954362	0.933943	0.944032
KNN	0.112972	0.353562	0.952780	0.962968	0.952783	0.957827
neural network	9.088517	0.006925	0.969879	0.978723	0.967605	0.973112
SVM_linear	1.647538	0.053979	0.927726	0.945592	0.926268	0.935779
SVM_poly	1.048257	0.074207	0.949254	0.968816	0.941779	0.955083
SVM_rbf	1.341540	0.103329	0.952149	0.968815	0.946580	0.957543
SVM_sigmoid	1.344607	0.109696	0.827498	0.846515	0.844311	0.845305
gradient boosting	0.891888	0.005298	0.948621	0.962481	0.946234	0.954260

Fig. 17. Performance by Paper [14].

In Fig. 18 shows the performance done by the previous paper [6]. The highest performance of the paper [6] is done by using random forest [19] with wrapper based feature selection which is about 97.3.

	Measures	Without features selection	With features selection		
			Wrapper	*PCA*	*IG*
BPNN	*TPR*	0.966	**0.971**	0.961	0.969
	TNR	0.963	**0.969**	0.958	0.967
	GM	0.964	**0.970**	0.959	0.968
RBFN	*TPR*	0.919	**0.931**	0.903	0.919
	TNR	0.917	**0.926**	0.902	0.917
	GM	0.918	**0.928**	0.902	0.918
NB	*TPR*	**0.929**	0.927	0.911	**0.929**
	TNR	**0.924**	0.922	0.907	**0.924**
	GM	**0.926**	0.924	0.909	**0.926**
SVM	*TPR*	0.944	**0.964**	0.946	0.944
	TNR	0.94	**0.962**	0.942	0.94
	GM	0.942	**0.963**	0.944	0.942
C4.5	*TPR*	0.958	**0.961**	0.952	0.959
	TNR	0.955	**0.958**	0.949	0.956
	GM	0.956	**0.959**	0.950	0.957
kNN	*TPR*	**0.971**	**0.971**	0.969	**0.971**
	TNR	0.969	**0.97**	0.966	0.969
	GM	**0.970**	**0.970**	0.967	**0.970**
RF	*TPR*	0.972	**0.973**	0.969	**0.973**
	TNR	0.969	**0.97**	0.967	**0.97**
	GM	0.970	**0.971**	0.968	**0.971**

Fig. 18. Performance of Paper [6].

The performance of our proposed system is given in Table I.

TABLE I. PERFORMANCE OF OUR PROPOSED SYSTEM

Model	Features	Accuracy
Random Forest	Wrapper	97.377
Random Forest	No Feature Selection	97.40
Random Forest	Embedded	97.1058
Random Forest	Correlation Coefficient	97.47
Random Forest	Information Gain	96.65
Random Forest	Chi Square	97.467
Light Gradient Boosted	Wrapper	97.36
Light Gradient Boosted	No Feature Selection	97.40
Light Gradient Boosted	Embedded	97.07
Light Gradient Boosted	Correlation Coefficient	97.40
Light Gradient Boosted	Information Gain	96.77
Light Gradient Boosted	Chi Square	95.99
Category Boosting	Embedded	96.29
Category Boosting	Correlation Coefficient	97.13
Category Boosting	Information Gain	96.77
Category Boosting	Chi Square	95.9

In our proposed system, the performance are given in Fig. 19. Here the overall performance is better than the previous references. We can see that Random forest model with wrapper feature selection shows the accuracy of 97.377%, with no feature selection and also with correlation shows slightly better result as 97.40%, 97.47%. Light Gradient boosted and Catboost also have better performance with the features selection methods. But most accuracy are shown by using Random Forest with Coefficient based feature selection which is approximately 97.47% RF with Coefficient feature selection for our proposed system.

Then we have used react native to get the real life application using our proposed method. we have used flask server as backened. Then in the traditional approach we have extract a url which will be collected from user then extract it into our desired features. Next pickle file of ml is connected with that. Lastly, frontened is designed which can get data from user and give them a result using our machine learning approach.

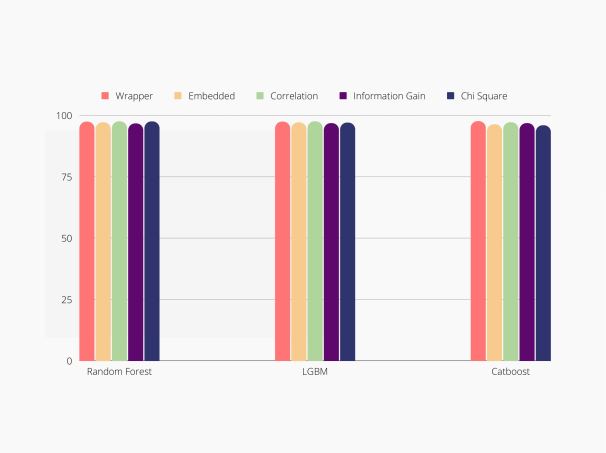


Fig. 19. Performance Measurement.

Fig. 20 shows the home page of the application. The input is given by the user and then it will be extracted and test whether its real url or fake one. The results of the input url can be fake or real which is given side by side. Fig. 21 showing that input url is malicious as well as fake one. Fig. 22 shows the input url is real one.

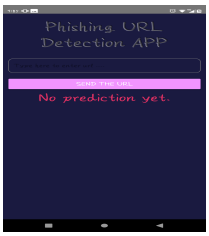


Fig. 20. Home Page.

Fig. 21. Fake Url.

V. Conclusion and Future Work

We have precisely improved the performance using machine learning techniques with feature extraction methods comparing with the traditional approaches. we have found above 97.47% of performance using these methods. We have used both Machine Learning and feature extraction techniques to create the models. We have trained and validate from same distributed data to create the model and evaluated it from different distributed of test data. After the evaluation the best models are Random Forest combining with Coefficient feature extraction. Next we have used our machine learning model to create an application by using react native.The app can classify whether the url is fake or real by using our machine learning model.

Due to some hardware limitations and the limited amount of time, we couldn't use more existing methodologies like deep learning, neural network etc.We had to experiment continuously and carefully to build the models and use every possible methods. We couldn't apply all the features for our existing application. And the app is hosted on local server. So in future it's structure can be better by using all features We have already got from our machine learning model and also can improve the user interface. Different types of embedding techniques also can be used and the existing methods like deep learning, tuning, data cleaning can better its accuracy.

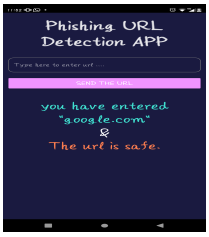


Fig. 22. Safe Url.

References

[1] S. Das, A. Kim, Z. Tingle, and C. Nippert-Eng, "All about phishing: Exploring user research through a systematic literature review," *arXiv preprint arXiv:1908.05897*, 2019.

[2] Z. A. Wen, Z. Lin, R. Chen, and E. Andersen, "What. hack: engaging anti-phishing training through a role-playing phishing simulation game," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[3] J.-N. Tioh, M. Mina, and D. W. Jacobson, "Cyber security training a survey of serious games in cyber security," in *2017 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2017, pp. 1–5.

[4] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge, "Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish," in *Proceedings of the 3rd symposium on Usable privacy and security*, 2007, pp. 88–99.

[5] G. Canova, M. Volkamer, C. Bergmann, and R. Borza, "Nophish: an anti-phishing education app," in *International workshop on security and trust management*. Springer, 2014, pp. 188–192.

[6] W. Ali, "Phishing website detection based on supervised machine learning with wrapper features selection," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 9, 2017.

[7] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[8] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[9] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms," *Machine learning*, vol. 40, no. 3, pp. 203–228, 2000.

[10] L. Rokach, "Ensemble-based classifiers," *Artificial intelligence review*, vol. 33, no. 1, pp. 1–39, 2010.

[11] M. A. U. H. Tahir, S. Asghar, A. Zafar, and S. Gillani, "A hybrid model to detect phishing-sites using supervised learning algorithms," in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2016, pp. 1126–1133.

[12] L. Wu, X. Du, and J. Wu, "Mobifish: A lightweight anti-phishing scheme for mobile phones," in *2014 23rd International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 2014, pp. 1–8.

[13] A. Zamir, H. U. Khan, T. Iqbal, N. Yousaf, F. Aslam, A. Anjum, and M. Hamdani, "Phishing web site detection using diverse machine learning algorithms," *The Electronic Library*, vol. 38, no. 1, pp. 65–80, 2020.

[14] V. Shahrivari, M. M. Darabi, and M. Izadi, "Phishing detection using machine learning techniques," *arXiv preprint arXiv:2009.11116*, 2020.

[15] R. Mohammad, L. McCluskey, and F. Thabtah, "Uci machine learning repository: Phishing websites data set," *archive. ics. uci. edu*, 2015.

[16] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," *Proceedings of international journal of advanced research in computer science and software engineering*, vol. 3, no. 6, 2013.

[17] S. Srivastava and R. Kumar, "Design and implementation of disaster management application using react-native."

[18] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from urls," *Expert Systems with Applications*, vol. 117, pp. 345–357, 2019.

[19] G. Biau, "Analysis of a random forests model," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1063–1095, 2012.

A Prototype Implementation of a CUDA-based Customized Rasterizer

Nakhoon Baek
School of Computer Science and Engineering
Kyungpook National University
Daehak-ro 80, Daegu 41566, Korea

***Abstract*—In these days, we have high-performance massively parallel computing devices, as well as high-performance 3D graphics rendering devices. In this paper, we show a prototype implementation of a full-software 3D rasterizer system, based on the CUDA parallel architecture. While most of previous CUDA-based software rasterizer implementations focused on the triangle primitives, our system includes more 3D primitives, and extra 2D primitives, to fully support 3D graphics library features. Currently, our system is at its prototype implementation stage, and it shows successful results with 3D primitive handling and also character output features. Our design and implementation details are presented. More optimizations and fine tunes will be followed in near future.**

Keywords*—3D rasterization; CUDA implementation; OpenGL emulation

I. Introduction

Web3D applications are designed to fully display and navigate web sites using 3D graphics features. Fundamentally, the Web3D applications need 3D rasterization process, either on the web-browser side, and/or on the web-server side. In this paper, we present a prototype implementation of 3D rasterizer, based on massively parallel processing features, as a framework for 3D web and associated application domains.

Massively parallel computing features are now widely available in many areas of computer science and engineering. From the 3D graphics rendering point of view, the 3D rendering pipelines are naturally developed to use massively parallel processing features. Additionally, we can also use the massively parallel computing features, especially with CUDA (compute unified device architecture) [1] and OpenCL (open computing language) [2].

Since these massively parallel pipelines, the 3D graphics pipeline and the high-performance computing pipeline, have many common characteristics, there have been several works to integrate these two different pipelines into a single one. More precisely, they tried to implement the 3D graphics rendering pipeline, on the existing parallel computing pipeline [3].

In previous works, they focused on the feasibility test, and most of them provides mainly the triangle rasterization process, with massively parallel computing libraries. In contrast, we aim at the full-scale 3D graphics rendering library implementation. For example, to provide the full features of the OpenGL (open graphics library) system, we need much more rather than the triangle rasterizer. At this time, we have a prototype implementation, which shows the possibility of the CUDA-based rasterizer, with several 3D graphics primitives and also extra 2D graphics primitives. It is the distinguished point of our work, in comparison to the previous works.

We start from presenting the previous works in Section II. Our motivation and overall design of the 3D rendering system based on CUDA will be presented in Section III. Implementation details and results from the prototype implementation will be followed in Section IV.

II. Previous Works

In 1990's, the programmable graphics pipeline has been introduced [4]. Rapidly, the GPU (graphics processing unit) became a computing device, with the concepts of GPGPU (general purpose GPU). In 2000's, the massively parallel computing devices including CUDA [1] and OpenCL [2] are available.

Since the programmable graphics pipeline and the massively parallel computing pipeline have common features, there have been several trials to implement the graphics processing features on the massively parallel processing devices. Some of them are summarized in the followings.

Mesa 3D graphics library [5] was originally implemented with CPU computing powers. Mesa is actually started as a re-implementation of widely-used 3D graphics libraries, including OpenGL [6] and Vulkan [7]. In its components, Mesa provides a full-software implementation of rasterizers, called "swrast". This implementation enables the 3D graphics rendering and 3D graphics shader features on CPUs. However, this CPU-based implementation is very slow, as we can expect, and thus, used only for limited purposes. Since this implementation was already available in 1990's, it actually affected its following implementations of software 3D graphics rasterizers.

Intel Larrabee project [8] was actually a hardware architecture, while it also aimed to provide an efficient software implementation of the 3D graphics rendering features. From the parallel processing point of view, the Larrabee project implements a binned renderer to increase the parallel processing features, and to reduce the memory bandwidth. Unfortunately, the Larrabee project was cancelled, and later rearranged to make high-performance computing processors.

FreePipe [9] is a fully-programmable 3D graphics pipeline, implemented in CUDA programming library. It developed some special features for the efficient rendering, even in a single pass rendering, with CUDA atomic operations. It shows good performance for small-size objects, while the

performance drops rapidly for large-scale and/or large-size objects, mainly due to the CUDA atomic operation behaviors.

CUDARaster [10] and its followers [11], [12], [13] are also software 3D graphics rendering pipeline implementations, with CUDA. CUDARaster was implemented for a specific CUDA model of Fermi, and it also uses some assembly-level codes, for optimization purpose. Unfortunately, it cannot be executed on the new CUDA architectures, since it was highly tuned and dependent on the old CUDA architecture.

The cuRE [14] is another rasterizer implementation to resolve the drawbacks of the previous implementations. This new rasterizer architecture can be executed on various modern CUDA architectures. It also shows several modifications, including direct wireframe rendering, programmable blending, and others.

Although we have some rasterizer implementations, especially based on the CUDA parallel processing architecture, our work aims to finally implement the full-scale 3D graphics system. Thus, we will include more 3D graphics primitives, and also 2D graphics primitives. The design and implementation of our system will be presented in the following sections.

III. Overall Design

In the case of OpenGL, they need at least the following 3D primitives:

- Points: A set of 3D points can be displayed. Additionally, they can set the radius of the point, or equivalently, the point size. The initial point size of 1 means a single pixel point, while we can also specify more big size points, which will be displayed as circles or rectangles on the screen.
- Line segments: A pair of 3D points can specify a 3D line segment. A sequence of line segments can also be displayed. In most cases, they use the line width of 1, to show one pixel wide line segments. With larger line widths, we can display thick line segments.
- Triangles: A set of three 3D points can define a triangle. A sequence of triangles are also possible, with triangle strips and triangle fans. Those triangles are usually filled with specified colors, or texture images.

In the previous works, they concentrated on the triangle primitives. In fact, the CUDA-based implementation of the triangle primitive is sufficiently difficult work, to be optimized and finely tuned. Also, we need to consider that most of 3D graphics scenes are constructed with triangles, since modern computer graphics object models are mostly based on the 3D triangle mesh models.

For a full-scale 3D graphics library implementation, we naturally need all of these 3D output primitives: points, line segments, and triangles. Additionally, for practical reasons, some 2D primitives are also needed to full-scale implementations. As an example, the resetting or updating of 2D rectangular areas in the framebuffer areas and/or in the texture image areas are needed frequently, even for the 3D graphics libraries.

Mostly required 2D operations are actually *bit-blt* (bit block transfer) operations, and can be summarized as follows:

- rectangular fill: The given rectangular 2D framebuffer (or image) area will be updated with the given colors (or numerical values). This operation can be used for the clearing of the whole or any partial framebuffer area.
- pixmap bit-blt: A pixmap means a 2D array layout of pixel values. A colorful image can be the typical cases. The image will be transferred (or more precisely, copied) to the specified rectangular area in the framebuffer or in the texture image area.
- grayscale bit-blt: A grayscale image can also be transferred to the framebuffer area.
- bitmap bit-blt: A bitmap represents a black/white image, through representing a black/white pixel with a single bit. This format is frequently used for bitmap fonts. Through transferring the bitmaps on the screen, we can display characters on the screen for extra information display.

For the overall design of the system, the 3D output pipeline will be maintained as the main stream pipeline. Fig. 1 shows the full 3D graphics pipeline of OpenGL 2.0 and OpenGL ES 2.0, which supports the vertex shaders and fragment shaders. It is used as the start point of our implementation. Our current prototype implementation focuses on the primitive assembly and rasterizer module.

The essential role of the "primitive assembly and rasterizer" module is converting the given 3D coordinate specifications to a set of 2D pixels, which are targets to be updated. The vertex shader and the fragment shader can be regarded as the preprocessing and post-processing to this module.

At this time, our CUDA-based implementation is based on the commercial CUDA-capable graphics cards. Thus, our prototype implementation is realized as a set of CUDA kernel programs, as shown in Fig. 2.

User inputs and rendering operations are provided through the C/C++ API function calls, to prepare the 3D graphics data in the CPU memory area. The 3D graphics data will be copied to the CUDA memory area, similar to the typical CUDA programs.

A big-size CUDA memory area is dedicated to the "logical framebuffer", which act as the real framebuffer, but cannot be displayed directly on the screen. Instead, the "logical framebuffer" is shared as an OpenGL texture image, and an independent OpenGL program is executed to simply display the logical framebuffer texture image on the screen.

For embedded systems, we can customize the current implementation, with new hardware display circuit supports, as shown in Fig. 3. With customized display logic implementations, the "logical framebuffer" can act as the real physical framebuffer. In this case, the display speed will be much more enhanced. This customized display circuit support will be the future works. In the next section, we will explain our current CUDA kernel implementations.

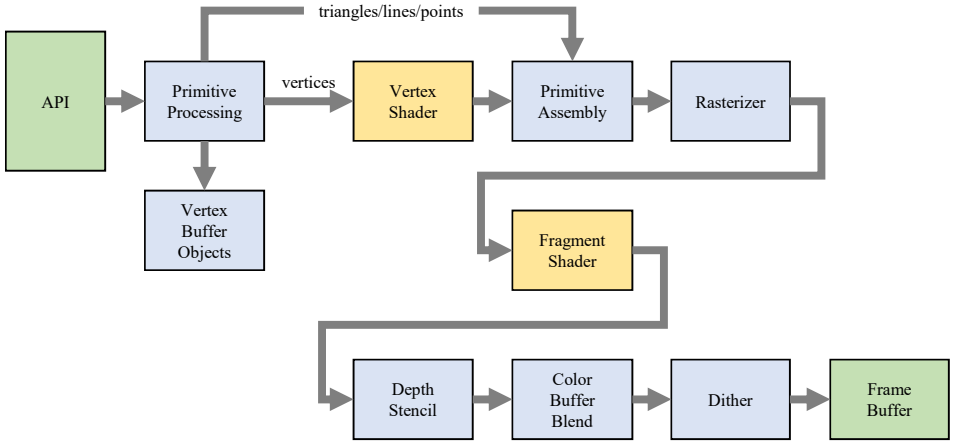


Fig. 1. A Typical 3D Graphics Rendering Pipeline.

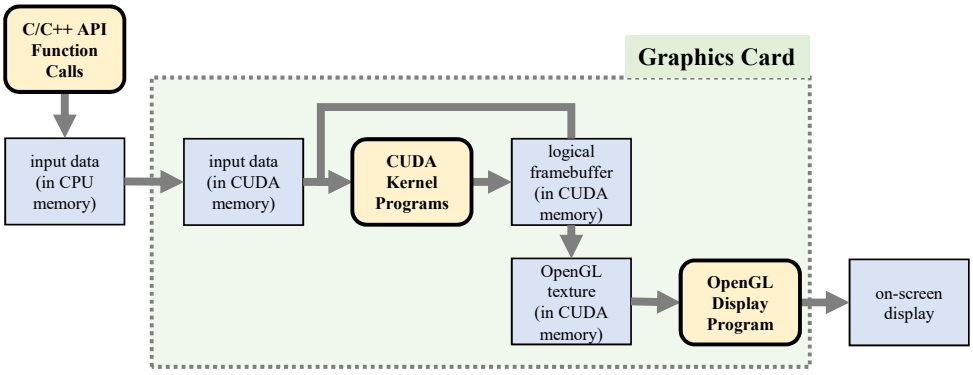


Fig. 2. Our Current CUDA-based Implementation Layout.

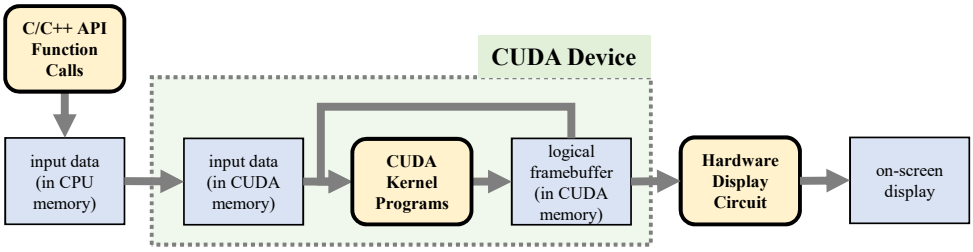


Fig. 3. Another Possible CUDA-based Implementation Layout.

IV. Implementation

The core of our implementation is a set of CUDA kernel programs, whose layouts are based on the rectangular division of the screen. It is actually typical approaches used in most previous works. As shown in Fig. 4, the whole screen (as an example, 1,280 by 1,024 pixels) is divided into a set of rectangular tiles. Each tile consists of 32 by 32 pixels, or equivalent, 1,024 pixels.

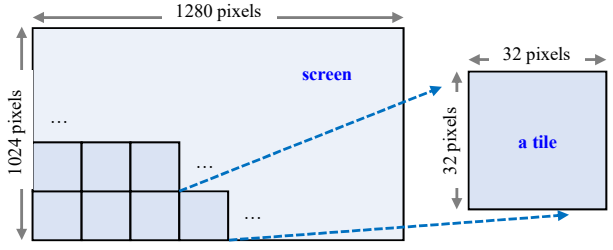


Fig. 4. The Screen Layout with Tiles.

From the CUDA programmer's point of view, it is convenient to allocate a single thread for a pixel on the screen. Thus, as shown in Fig. 5, we use a 2D thread block of 32-by-32 thread layout. This 1,024 threads are actually the current CUDA limits to the maximum number of threads in a single thread block. Then, to make the whole 1,280-by-1,024 threads,

we use 40-by-32 blocks for the CUDA grid. Thus, the whole grid corresponds to the whole screen. The grid is divided into 40-by-32 thread blocks, while the screen divided into that numbers of tiles. And, the 32-by-32 thread block corresponds to the thread block.

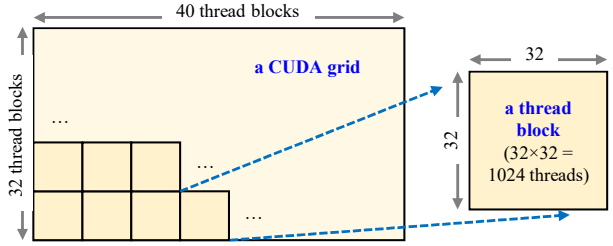
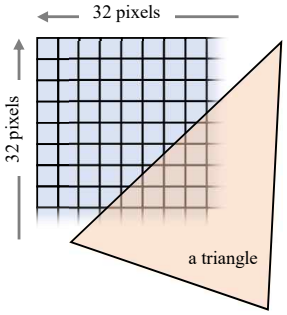
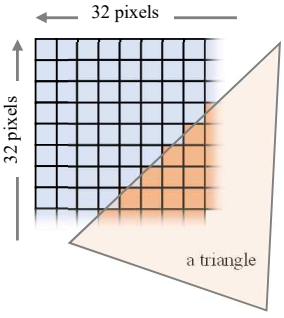


Fig. 5. The CUDA Grid Layout with Thread Blocks.

This thread block layout has some benefits. In the case of triangles, the pixels, or equivalently, the CUDA threads can decide whether they are located in the interior of a given triangle or not, in a massively parallel way. Each thread will calculate the signed areas of some configurations, from the given window coordinates of the vertices. Only the interior threads will turn on their corresponding pixels, to display the given triangle on the screen, as shown in Fig. 6.



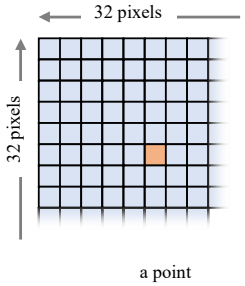
(a) A Triangle Overlaid on the Tile



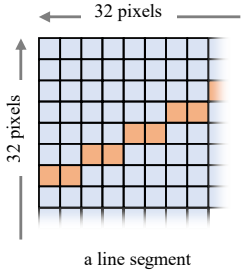
(b) The Triangle Rasterization Result

Fig. 6. A Tile-Based Rasterization Example for a Triangle.

In the case of points and line segments, the tile based approach can be inefficient, especially for the single pixel points and the single pixel wide line segments. For a single pixel point, the thread block should launch totally 1,024 threads, due to the CUDA kernel launch mechanism and our thread block configurations. Then, only one thread will turn on the pixel, while others all should discard their processing, as shown in Fig. 7(a). Similarly, a single pixel wide line segments, at most 32 pixels will be turned on, even though initiating totally 1,024 threads, as shown in Fig. 7(b).



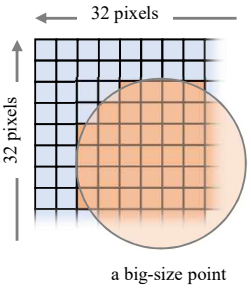
(a) A Single-Pixel Point Case



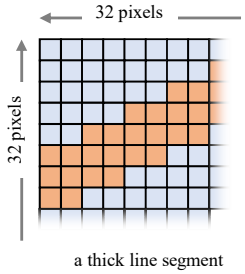
(b) A Line Segment with One Pixel Width

Fig. 7. A Tile-Based Rasterization Example for a Point and a Line Segment.

In contrast, the tile based approach can efficiently works with big-size points, and think line segments. As shown in Fig. 8(a), the big-size points are typically implemented as circles, and the threads in the tile can check whether they belongs to the interior of the circle or not, similar to the triangle cases. Thick line segments can also be handled efficiently, as shown in Fig. 8(b). The effective threads can check their corresponding conditions in a massively parallel manner.



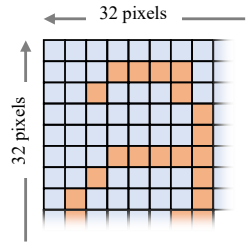
(a) A Big-Size Point Case



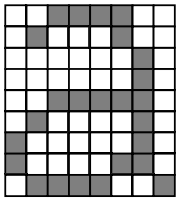
(b) A Thick Line Segment Case

Fig. 8. Another Tile-Based Rasterization Example.

The tile-based approach can work for most of bit-blt operations. For a given rectangular region, the threads will get the corresponding pixel information, and then update their own pixels, to get the final result. As a direct application of these bit-blt operations, we added character display features to our implementation. In this case, the true type fonts are pre-processed to get the character font information and the grayscale image of each character [15], as shown in Fig. 9(b). Our thread blocks will process the character font information, and finally show the character on the screen, as shown in Fig. 9(a).



(a) A Character Display Example



(b) A Grayscale Image from True Type Fonts

Fig. 9. A Tile-Based Rasterization for Character Display.

At this time, our prototype implementation shows the basic rasterization CUDA kernels are working well. As an example, Fig. 10 shows the screen shot of the triangle rasterization result, from our CUDA-based rasterizer implementation. It shows the correct display of the triangle coordinates, as specified in the input vertex specifications.

Additionally, the barycentric interpolation of interior points are also demonstrated. We specified different vertex colors at each vertex of the triangles. The interior pixels have the interpolated colors, according to the barycentric interpolation, specified in the OpenGL specification [6].

Unlike the previous works, our CUDA-based rasterizer implementation supports more output primitives, in addition to the triangle primitives. Fig. 11 shows a demonstration of 3D points, from our prototype implementation. It shows the circular points, as expected.

As another distinct example, we implemented the text output routines, with underlying bit-blt primitive support. Our CUDA kernels support bit-blt operations, and we use some grayscale or bitmap images of the true type fonts, with the free true type font library [15]. The images are blended into the screen, to make smooth font display results. Fig. 12 shows an example screen shot of our font rendering result, with more than 100 text output results, each of which specify random text colors and a complete sentence to be displayed. It shows that our CUDA-based implementation has some

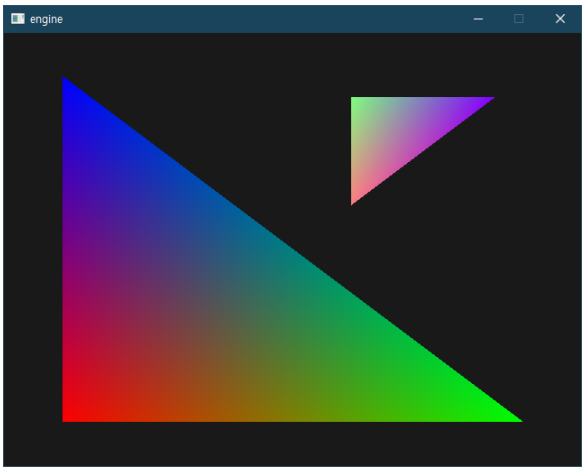


Fig. 10. A Screen Shot of Triangle Rasterization, from our CUDA-Based Rasterizer Implementation.

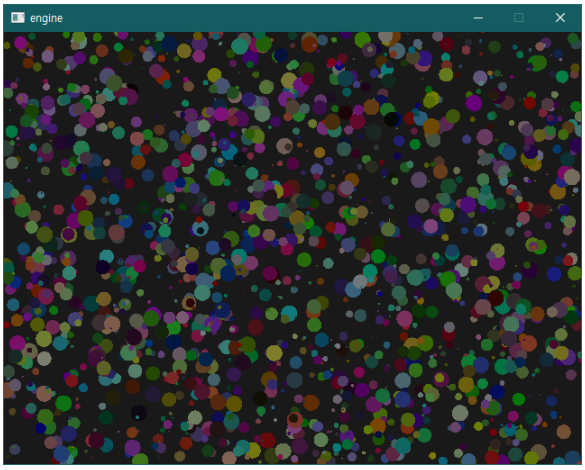


Fig. 11. A Screen Shot of Point Rasterization, from our CUDA-Based Rasterizer Implementation.

distinguished points, in comparison to the previous rasterizer implementations.

V. Conclusion

Our motivation was implementing the 3D graphics rendering pipeline on the massively parallel computing pipeline. In this case, we can make a full-software implementation of the graphics rendering features. To realize this goal, we started to implement common 3D rendering features on the CUDA architecture.

Since we aimed to get a full-scale implementation of typical 3D graphics library, we selected several 3D graphics primitives including points, line segments and triangles. Additional image handling operations are also needed, and we added some 2D pixel level primitives. Currently, our prototype implementation shows those primitives are working well. More fine tunings and optimizations should be followed, and they will be our near future works.

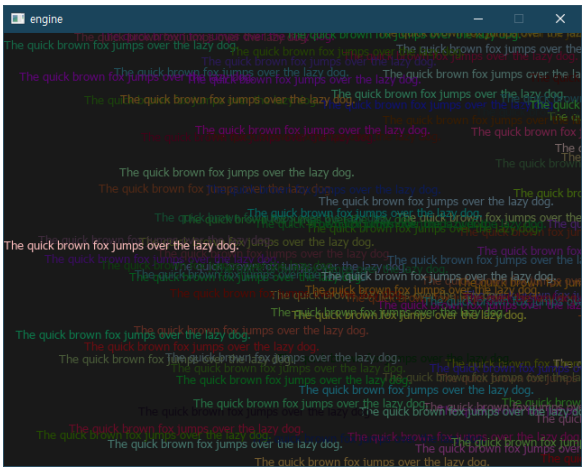


Fig. 12. A Screen Shot of True-Type Font Rasterization, from our CUDA-Based Rasterizer Implementation.

ACKNOWLEDGMENT

This work has supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grand No.NRF-2019R1I1A3A01061310).

This study was supported by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (4199990214394).

REFERENCES

[1] NVIDIA, *CUDA Toolkit Documentation, version 11.7.0*. NVIDIA, 2022.

[2] Khronos OpenCL Working Group, *The OpenCL Specification, version 3.0*. Khronos Group, 2022.

[3] J. E. Stone, D. Gohara, and G. Shi, "OpenCL: A parallel programming standard for heterogene-ous computing systems," *Computing in science & engineering*, vol. 12, no. 66, 2010.

[4] D. Kirk, "NVIDIA CUDA software and GPU parallel computing architecture," in *Proc. of the 6th Int'l Symp on Memory Management (ISMM '07)*. ACM, 2007, pp. 103–104.

[5] Mesa Team, *The Mesa 3D Graphics Library*, retrieved in July 2022. [Online]. Available: http://www.mesa3d.org/

[6] M. Segal and K. Akeley, *The OpenGL Graphics System: A Specification, version 4.6*. Khronos Group, 2019.

[7] The Khronos Vulkan working group, *Vulkan - A Specification, version 1.3.223*. Khronos Group, 2022.

[8] L. Seiler *et al.*, "Larrabee: A many-core x86 architecture for visual computing," *IEEE Micro*, vol. 29, pp. 10–21, 2009.

[9] F. Liu, M. C. Huang, X. H. Liu, and E. H. Wu, "Freepipe: A programmable parallel rendering architecture for efficient multi-fragment effects," in *Proc. of the 2010 ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3D '10)*. ACM, 2020, pp. 75–82.

[10] S. Laine and T. Karras, "High-performance software rasterization on gpus," in *Proc. of the ACM SIGGRAPH Symp on High Performance Graphics (HPG '11)*. ACM, 2011, pp. 79–88.

[11] Y. C. Kwon and N. Baek, "A cuda-based implementation of opengl-compatible rasterization library prototype," in *Proc. of the 29th Annual ACM Symp on Applied Computing (SAC '14)*. ACM, 2014, pp. 1747–1748.

[12] N. Baek and K. Kim, "Design and implementation of opengl sc 2.0 rendering pipeline," *Cluster Computing*, vol. 22, pp. 931–936, 2019.

[13] M. Kim and N. Baek, "A 3d graphics rendering pipeline implementation based on the OpenCL massively parallel processing," *Journal of Supercomputing*, vol. 77, pp. 7351–7367, 2021.

[14] M. Kenzel, B. Kerbl, D. Schmalstieg, and M. Steinberger, "A high-performance software graphics pipeline architecture for the GPU," *ACM Trans. Graph.*, vol. 37, pp. 140:1–140:15, 2018.

[15] FreeType, *The FreeType project*, retrieved in July 2022. [Online]. Available: http://www.freetype.org/

Mobile App Design: Logging and Diagnostics of Respiratory Diseases

Diana Cecilia Chávez Cañari[1], Ángel Vicente Garcia Obispo[2], Jose Luis Herrera Salazar[3], Laberiano Andrade-Arenas[4], Michael Cabanillas-Carbonell[5]
Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú[1,2,3,4]
Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú[5]

***Abstract*—Over the years, a wide variety of respiratory diseases have caused a high mortality rate throughout the world. This was again observed with the appearance of the pandemic, COVID-19. In addition, the most affected are people living in extreme poverty. The objective is design a mobile health application for the registration and diagnosis of respiratory diseases. For this, the RUP methodology was applied, because it easily adapts to various types of projects. Its use, together with the UML process development software, allows the analysis, implementation and documentation of object oriented systems. For validation, a user survey was carried out and the questionnaire was based on the dimensions of functionality, efficiency, effectiveness and satisfaction. Obtaining as a result a positive qualification to the design of the application and its acceptance due to the reduction in the time to obtain the diagnosis. In conclusion, a mobile health application design was successfully carried out so that patients can register and have the diagnosis of respiratory diseases from the comfort of their home.**

***Keywords*—Mobile app; Covid-19; diagnosis; respiratory diseases; RUP methodology**

I. Introduction

In the world there are a large number of diseases of varying degrees of danger that affect the respiratory system. From 2020 to the present, COVID-19 caused by the SARS-CoV-2 [1], virus is characterized by symptoms similar to those of a common cold. As it progresses, it causes multi-organ damage, including respiratory distress [2]. Therefore, this pandemic has generated a high rate of morbidity and mortality worldwide. The people most likely to have a fatal prognosis are those who present pre-existing diseases to contagion, so they are classified as a high-risk group. The most vulnerable people are the poorest people [3].

Another disease that strongly affected Latin America is influenza. This was observed in the research [4], where the author compared information from 10 countries (Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica, Panama, Ecuador, Brazil, Argentina and Chile). The data analyzed were influenza A subtypes H1N1, H1N1pdm09, and H3N2, influenza B subtypes Victoria and Yamagata, and non-subtypes of both types. The results showed that of the 37,087 cases reported between the years 2004-2012, the most predominant was type A influenza.

On the other hand, the investigation [5], was carried out in Peru, between the years 2011 and 2016 due to the large number of deaths caused by acute respiratory infections (ARI). The purpose of this study was to make known which were the most affected departments. For this, geographic information systems were used. From there, information was obtained on the existing conditions in those areas. The findings showed the various factors that caused the disease in Peruvian children under 5 years of age.

Thus, currently going through the global pandemic, COVID-19, people have become more aware of health care [6]. This situation allowed mobile healthcare applications to expand and evolve faster and stronger. The development of smart mobile health applications allow improving the effectiveness and efficiency of various processes. The use of monitoring features, appointment booking scheduling, self-diagnosis, emergency care and home visit received a positive response from users. Given these results, the RUP methodology was used in the software development process because it provides techniques that team members must follow in order to increase their productivity and generate a high-quality product.

In this context, the present work objective to design the prototype of a mobile health application that presents characteristics of remote registration and diagnosis through a questionnaire of symptoms. Being these relevant factors to avoid the prolonged time of exposure to various pathogens that further complicate their clinical picture. Therefore, its importance lies in improving the effectiveness and efficiency of the registration control processes and patient care [7].

Finally, the structure of the work is broken down into six sections. Section II explains the review of the literature, section III the methodology, section IV the results, section V discussions, and finally section VI the conclusions and future work.

II. Literature Review

This section focuses on analyzing the different investigations related to this research work, finding its results and conclusions.

A. Background

COVID-19 appeared in Wuhan, China and is caused by the SARS-CoV-2 virus. This strain that infects humans became a pandemic due to its easy transmission. Reason that caused a high rate of sick and dead around the world. The studies carried out showed that the greatest number of deceased were

elderly people and those who had pre-existing diseases. Little by little, the specialists were documenting information about this lethal disease and sharing it with the rest of the countries in order to find a solution. In addition, the most reputable web platforms around the world were very important in order to obtain real-time statistics on reported cases [8]

The rapid increase in cases of the global COVID-19 pandemic was due to the fact that doctors in this country initially had great difficulty in identifying infected people. This highly contagious disease is characterized by causing acute respiratory distress syndrome. Therefore, in case of presenting some other disease, the mortality rate increases [9].

In Peru, the MINSA website reports data about the population and Covid-19, such as the number of positive cases, numbers of hospitalized, deceased and vaccinated, number of people who underwent antigen and rapid tests. This information is found segmented by departments of the country and months, but general statistics are also found [10].

B. Related Work

The Rational Unified Process (RUP) [11] methodology can simplify the process of analysis and design of information systems, but of course, each method will have its advantages and disadvantages in certain situations and conditions. Using the RUP method you can accept changes to improve existing prototypes so that they can produce an acceptable system, and the changes that occur are considered as part of the development process itself.

This research [12] explains that mobile applications, by integrating medical records electronically into the health record system of hospitals, facilitate the management of medical treatments and interventions. The application made for the local hospital Sidi Said located in Meknès - Morocco resulted in a high rate of approval by patients by obtaining positive results in improving their health by monitoring their illnesses at home and by the patients. doctors, who with timely information could make better health decisions.

Today there are various techniques applied in desktop, web and mobile environments; which facilitate the human being to carry out a series of processes according to the author [13]. In this sense, the use of technologies for the diagnosis of respiratory diseases becomes a very favorable process when a chatbot is used together with a mobile application because it allows more accurate results.

Artificial intelligence (AI) occurs in different contexts such as industry, biology, computer science in order to give solutions. It has been widely demonstrated that in order to make this tool highly effective, the knowledge of expert professionals is required. According to [14] this entered information is contrasted with the data entered by those people who present symptoms of respiratory diseases. Using the information provided by experts increases the accuracy of the result that indicates the diagnosis of respiratory diseases presented by the individual.

The author's research [15] focuses on pneumonia; respiratory infection resulting in inflammation of the lungs. The causes of this respiratory disease can be attributed to viruses, bacteria or fungi. Rural people in developing countries have limited access to doctors, medical diagnostic centers, and hospitals. Therefore, for this article, a smartphone-based app for preliminary detection of pneumonia using X-ray images was designed and developed. The app was developed in Android Studio and incorporated the Tensor Flow library.

The mobile technology model proposed by the author [16], for online ambulatory health care information, uses a cloud platform. The model consisted of four phases: 1) The selection of structured data; 2) The integration and storage of data in a cloud database; 3) Real-time data testing using a data analytics service; 4) The results of the pharmacological consultation are displayed through a mobile application and the geolocation service can determine the closest pharmacies to the current location of the patient.

In summary, different research works were compared and it was observed that the authors focus on the development of mobile applications for the health sector with attributes such as functionality, quality and design. However, no reference is made to existing integration problems in the health system, which makes it difficult to obtain information from all health centers in the country.

III. Methodology

For the development of the application, the analysis of various methodologies was contemplated; According to [17] the development of good software depends on the use of adequate methodologies that allow compliance with existing standards for this type of project. This section explained the methodology, as well as the tools used in the development of a mobile application to improve the registration and diagnosis of patients with respiratory diseases in Lima-Peru. Its uses are due to the fact that they help in reaching objects in an agile way.

The methodology that served as a guide for the development of the project was the Rational Unified Process (RUP) methodology, according to [18] it allows adjusting various components and repeating phases of the cycle as many times as necessary until the software meets the requirements and objectives. Among the functions it offers are the assignment of tasks and responsibilities within the company to guarantee the production of high-quality software.

The development of the mobile application was carried out in Android Studio, according to [19], the integrated development environment (IDE) was introduced in the years of 2013 and is based on IntelliJ. What makes it a powerful code editor is that it provides built-in services and allows for a wide variety of customization options for Android app development.

SQL Server was used as a database manager as a knowledge base in which the user interacts in order to manage possible diagnoses based on the questions asked in the mobile application, according to [20] the importance of Database security is something that has to be considered in projects, giving as an example basic database security guidelines.

A. RUP

The methodology called Rational Unified Process (RUP), was applied with the purpose of creating high quality software. This agile methodology allows developing projects on a small

and large scale, since it adapts to the needs of different types of projects, especially in the use of each role. Additionally, it can also be used in projects that require reengineering [21].

1) RUP Life Cycle: Fig. 1 shows the life cycle of the RUP methodology where spiral development is implemented according to [22]. In the life cycle, tasks are performed in four stages or phases where a variable number of iterations occur. The first few iterations (during the Inception and Discovery phases) focus on understanding the problem and the technology, defining the scope of the project, eliminating critical risks, and establishing a baseline.

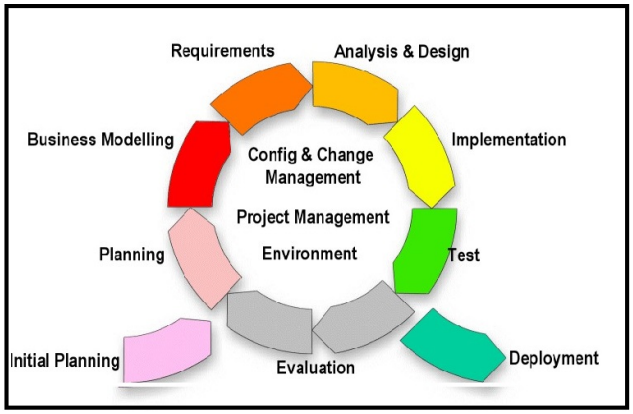


Fig. 1. RUP Life Cycle.

2) Phases of the RUP Methodology: The RUP methodology it is divided into four phases: Introduction, which defines the business case and scope of the project, identifies the most important use cases for the system and the stakeholders that interact with the system; in the development phase, an architectural prototype is built, which ensures minimizing risks and loss of information; construction phase to add functionality to the system; and the purpose of the transition phase is to deliver the system to the end users and make the appropriate adjustments or improvements, since failures often occur during testing [23].

3) Key Principles of the RUP: The RUP is based on 6 key principles; adapt to the process, maintain the balance of priorities, demonstrate value iteratively, allow collaboration between teams, focus on quality and raise the level of abstraction [24].

B. Tools and Programming Language to Develop the Prototype

This segment detailed the tools to be used to develop the prototype of the mobile application focused on improving the registration and diagnosis of respiratory diseases caused by Covid-19, as well as the programming languages to be used and the database.

1) Mockplus: The fabrication of the prototype was carried out in Mockplus [25], a tool that facilitates the creation of prototypes. It makes design faster, smarter and easier. It is integrated with more than 100 components with which any software prototype for mobile, web and desktop applications can be designed. Also, you can work online and offline on Windows PC.

2) Android Studio: Android Studio is an environment integrated development (IDE) that has a large number of methods and access to components that allows you to develop applications on Android. According to [26], this powerful code editor based on IntelliJ IDEA has a wide variety of development tools and offers various productivity functions.

3) Java: For the development of this application, the Java programming language because it is compatible with Android Studio. According to [27], this high-level programming language allows developers to write more robust, secure and stable code. It is widely used by programmers because it is a high performance, dynamic, simple and easy to understand language. In addition, it provides great advantages such as improving the efficiency of programming and the practicality of the software.

4) SQL Server: The Microsoft SQL Server tool (MSSQL) that allows you to store and manage the confidential data of both people and companies is regularly updated. This relational database manager is used by many users because it adapts correctly to various development platforms, such as .NET [28]. Additionally, according to [29] Android Studio with the Java language together with the SQL Server database. The integration of these technological tools allows the company's objectives to be achieved because it contributes to the improvement of processes where it is necessary to manage large volumes of data.

IV. Case Study

A. Development of the methodology

Based on what was mentioned above and the procedures that were detailed, an Android application based on registering and diagnosing people who present symptoms of respiratory diseases will be implemented.

1) Start Phase: This section defines the scope of the project with the clients, the risks associated with the project are identified, the plan of the phases and the subsequent iteration are drawn up, the software architecture is also detailed in a general way.

Definition of the Scope of the Project with the Clients:

To know the existing needs, the interview was used as a data collection instrument. The information obtained can be seen below:

- Patient: Let it be a mobile application to have quick access to registration and notifications in case of rescheduling of a medical appointment.
- Receptionist: Display the dates, hours and specialties of the doctors, taking into account the limit number of visits per shift and the shift changes made.
- Doctor: That the application be interactive with patients so that they can improve the diagnostic process.

Table I shows the methodologies and the platform used for the development of the project.

TABLE I. METHODOLOGY AND PLATFORM FOR DEVELOPMENT

N°	Description
1	Methodology learned in the university journey
2	Methodology used to date in various companies
3	Applied in small and large scale projects

2) Elaboration Phase: At this stage, define the requirements and is where a preliminary solution is designed, use cases are selected to define the underlying system architecture, and the first domain analysis is performed.

Definition and Determination of Requirements: After collecting the information from the business actors, the most relevant requirements for the development of the project are determined.

Table II mentions the functions that the system will present.

TABLE II. FUNTIONAL REQUIREMENTS

Code	Description
RF01	Enter the app
RF02	user management
RF03	Log management and diagnosis

Process Use Cases Table III lists the people who participate in the process related to the system access use case and the actions they should perform.

Table IV mentions the people who participate in the process related to the registry and diagnosis management use case and the actions that they must carry out.

Fig. 2 shows how the different parts of the system interact with each other in order to carry out a task, and the order in which the interactions are performed when executing a specific use case.

Table V shows the table based on the sequence model where the tasks to be performed by the system are mentioned.

3) Database Construction Phase: the basis of The aforementioned data was designed based on a rigorous analysis of the requirements that were obtained from the interviews with the different users, including doctors, patients and the receptionist. Observing the database from all perspectives will help avoid errors.

TABLE III. LOGIN USE CASE DIAGRAM

Use cases	Use cases
Actor	Doctor, Patient and Admission
Description	Each actor must enter the system, entering their username and corresponding password.
Preconditions	The staff requests user registration from the administrator.

TABLE IV. LOG MANAGEMENT AND DIAGNOSIS

Use cases	Log management and diagnosis
Actor	Doctor
Description	The medical actor will fill out the diagnostic questionnaire for respiratory diseases
Preconditions	The staff must have the role of doctor

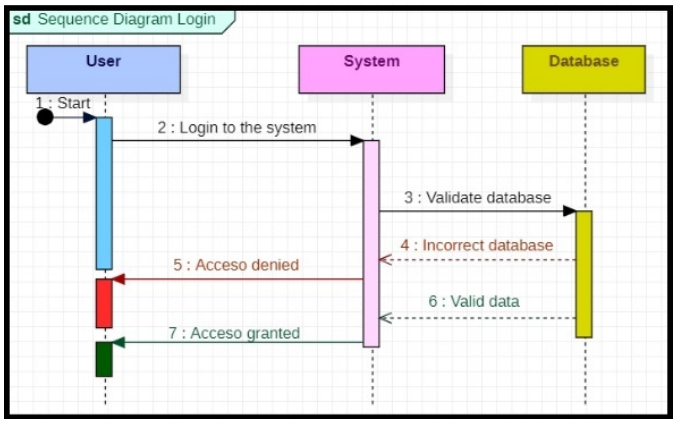


Fig. 2. Sequence Diagram Model.

TABLE V. SEQUENCE DIAGRAM MODEL

Code	Description
MDS01	Enter the app
MDS02	user management
MDS03	Log management and diagnosis
MDS04	care management

Fig. 3 shows the schema of the database in which the various tables and the relationships between different entities of the system are shown.

4) Construction Phase: The function of this phase is complete system functionality, clarify outstanding requirements, manage changes accordinglyo the evaluations made by the users, and improvements are made for the project.

In Fig. 4(a) you can see the login, which is where the user and password are validated to start the session and in Fig. 4(b) the home screen is shown in which there is a layer that is used to receive notifications, a section with information related to the respiratory system, a section for locating the closest hospitals to your location and an icon to display the menu.

In Fig. 5(a) the options menu is displayed to access the COVID-19 survey, survey history and profile interfaces and in Fig. 5(b) the Covid-19 prediction interface is observed where The symptoms presented by the patient are selected and saved.

In Fig. 6(a) is the interface where it is shown the graph of the probability of presenting a respiratory disease, below the recommendations and then options to communicate with a health personnel and in Fig. 6(b) the survey history details all the times the survey has been carried out, the disease detected, the date, state and probability percentage.

In Fig. 7(a) the user profile is presented, consisting of the photo, full name and role of the user, followed by icons of social networks with which it can be linked, under general vision options, settings and change of password and in Fig. 7(b) you can see the option to change the password displayed where you must enter the current password, the new one, repeat the new one and press the button to change the password for this to be done.

In Fig. 8 you can see the option to disconnect that allows

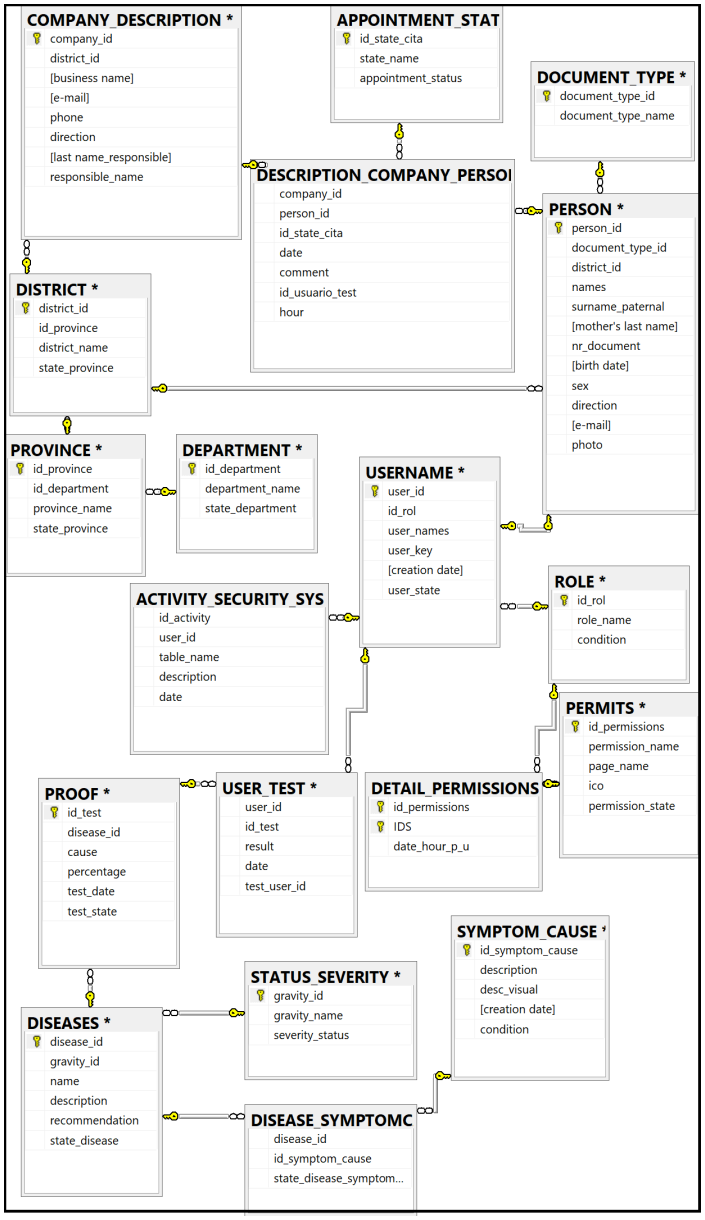


Fig. 3. Database.

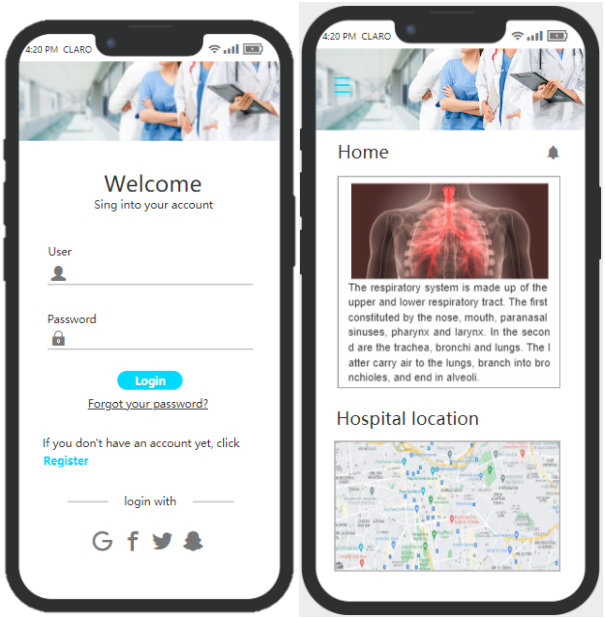
you to exit the session, which will redirect the user to the login view.

5) Transition Phase: In this phase, the purpose is ensure that the software is available to end users, correct errors and defects found in acceptance tests, train users and provide the necessary support.

V. Results

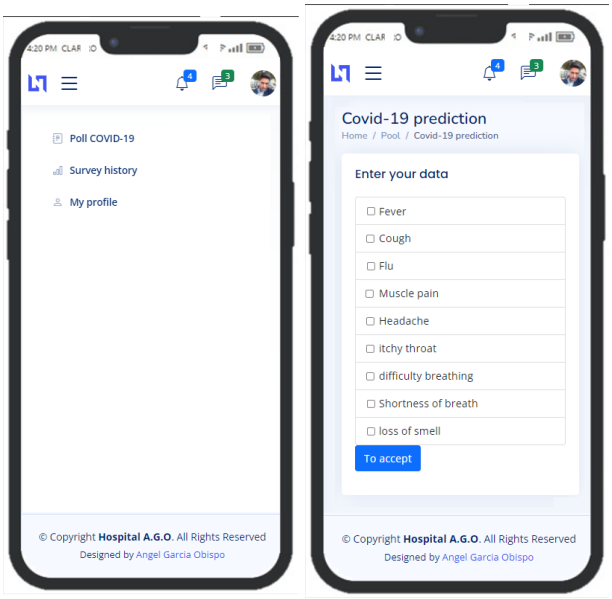
A. According to the Prototype

Fig. 4 presented the login and home modules. The first module allows entry to the session after verifying that the entered user has been previously registered; in addition, the password must correspond to said user; this is done for security reasons. Fig. 5 presented the menu of options offered by the application and the prediction of Covid-19. In the latter, the



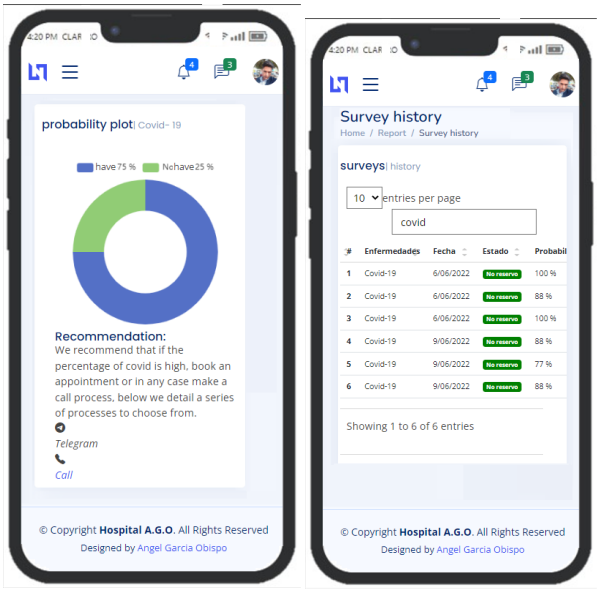
(a) HU-07 (b) HU-08

Fig. 4. Login and Home Prototype.



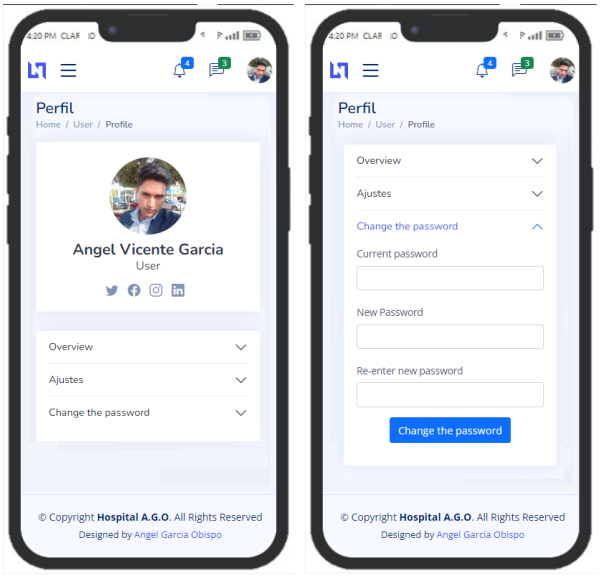
(a) HU-09 (b) HU-10

Fig. 5. Menu and Prediction of Covid-19.



(a) HU-11 (b) HU-12

Fig. 6. Probability of Covid-19 and Survey History.



(a) HU-13 (b) HU-14

Fig. 7. Profile and Password Change.

user selects the symptoms that they present and when pressing the To accept button, Fig. 6 will show the probability that a person has of have Covid-19 according to the information entered and the survey history where it is recorded each time the symptom survey has been carried out. On the other hand, Fig. 7 presented the profile and password change modules and finally in Fig. 8 the option to disconnect was presented, which allows to exit the session correctly.

1) The Results According to the Dimensions: What research method the survey technique was used to know the

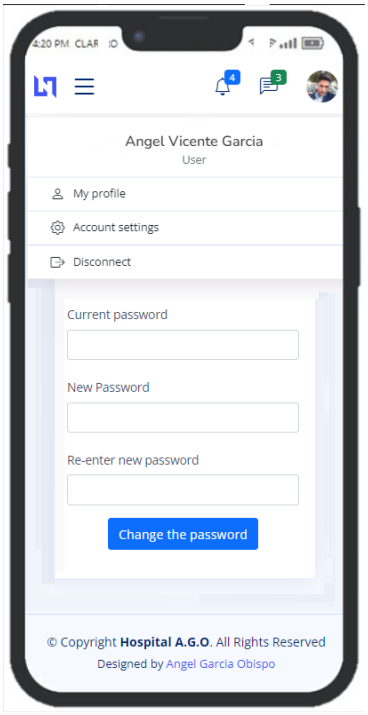


Fig. 8. Disconnect.

TABLE VI. ACCORDING TO THE SURVEY

Dimension	N°	Question
Functionality	P01	Do all the filling fields allow to enter the data correctly?
Functionality	P02	Do all the buttons work correctly?
Functionality	P03	Is the data saved correctly?
Functionality	P04	Does the application correctly show you the closest hospitals to the area where you are?
Functionality	P05	Does the signs and symptoms questionnaire allow multiple selections?
Effectiveness	P06	Do you perceive any slowness in the use of the application?
Effectiveness	P07	Does the result shown by the application correspond to the diagnosis given by your doctor?
Effectiveness	P08	Should the questionnaire be updated depending on whether new symptoms are found?
Efficiency	P09	¿La navegación es fácil e intuitiva?
Efficiency	P10	Does it take time to display the results or give an answer?
Efficiency	P11	Are all the functions of the application easily accessible?
Efficiency	P12	Did the application allow you to obtain a diagnosis quickly?
Satisfaction	P13	The results obtained that were evaluated with the doctor. Did I demonstrate the effectiveness of the application?
Satisfaction	P14	Do you feel that the application is user friendly?
Satisfaction	P15	Are you satisfied with the questionnaire?

opinion of users (patients, doctors and other users) about the application. To do this, 15 questions were asked that are grouped into four dimensions as shown in Table VI.

2) About the Survey: Upon completion of the user survey, it was possible to obtain 51 answers that allowed generating Table VII. It shows in detail the results obtained in the

response options yes, regularly and no presented by each of the questions asked. In Fig. 9, the results obtained in each of the dimensions are presented graphically according to the response options.

TABLE VII. RESULT OF EACH QUESTION IN PERCENTAGE

Dimension	Numero pregunta	Si	Regularmente	No
Satisfaction	P13	41%	39%	20%
	P14	100%	0%	0%
	P15	20%	78%	2%
Funcionalidad	P01	100%	0%	0%
	P02	100%	0%	0%
	P03	100%	0%	0%
	P04	41%	59%	0%
	P05	80%	0%	20%
Eficacia	P06	2%	59%	39%
	P07	59%	41%	0%
	P08	100%	0%	0%
Eficiencia	P10	2%	59%	39%
	P11	61%	39%	0%
	P12	22%	78%	0%
	P09	100%	0%	0%

The measurement of the scale was made based on 1 which denotes a high approval of the application, 2 denotes a medium approval of the application and 3 denotes a low approval of the application.

3) Phase of Assigning Scores to Securities: In this phase its purpose is to assign a range of values to the answers of the questions in order to know if the application meets the needs of the users. These will be in accordance with the assigned dimensions in order not to neglect the answers obtained. The score to be taken into consideration is found in Table VIII.

4) Assignment of Scale Indicators: In this phase its purpose is to define the percentage of the validity of each question with the sole purpose of obtaining a better validity of the program. Table IX serves to determine the scale in which the application is found.

5) About Data Analysis with SPSS: Next the graphs obtained from SPSS are detailed.

In Table X it is observed the Frequency percentage of each question and answer.

About the Chart Fig. 9 details the graph obtained based on the results obtained, favoring the implementation of the program, hoping that the survey obtained is the most appropriate to obtain the result.

TABLE VIII. ASSIGNMENT OF SCORES TO SECURITIES

Puntaje	Descripcion
1	Si
2	Regular
3	No

TABLE IX. ASSIGNMENT OF INDICATORS TO SCALES

Escala de porcentaje	Descripcion
1% al 30%	Bajo
31% al 60%	Regular
61% al 100%	Bueno

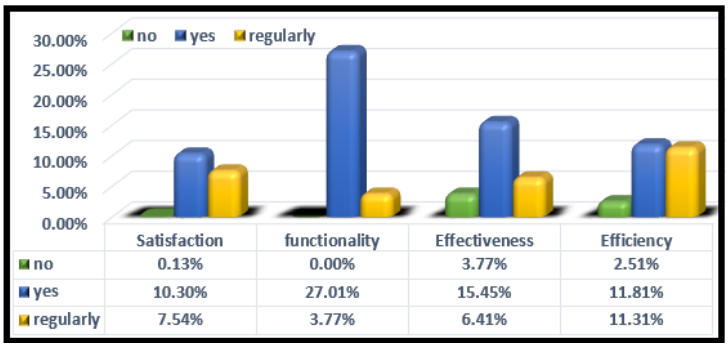


Fig. 9. Result in Percentage.

TABLE X. FREQUENCY PERCENTAGE OF EACH QUESTION AND ANSWER

Question	Response	Frecuencia	Porcentaje	% válido
P01	yes	11	100	100
P02	yes	11	100	100
P03	yes	11	100	100
P04	yes	5	45	45
	not	6	55	55
P05	yes	9	82	82
	regularity	2	18	18
P06	yes	2	18	18
	regularity	4	36	36
	not	5	45	45
P07	yes	5	45	45
	not	6	55	55
P08	yes	11	100	100
P09	yes	11	100	100
P10	yes	2	18	18
	regularity	4	36	36
	not	5	45	45
P11	yes	7	64	64
	not	4	36	36
P12	yes	4	36	36
	not	7	64	64
P13	yes	6	55	55
	not	5	45	45
P14	yes	11	100	100
P15	yes	2	18	18
	regularity	2	18	18
	not	7	64	64

B. About the Methodology

The selection of the methodology used in this research work was made after making a detailed comparison with each of them. Due to the qualities found and based on the nature of the project, the RUP methodology was chosen. The reasons why it stood out from the others is that despite being structured, it is also flexible and because it allows tests to be applied in the various phases of software development, which guarantees the delivery of a quality product.In the option of choosing between the Rup and Scum methodology, Rup was chosen since it was adapted to the research carried out; since Scrum works with sprint and finished product adapting to change, instead Rup allowed to work with the identification of functional requirements that are modeled through prototype design. It also allowed to make the documentation for each stage.

VI. DISCUSSIONS

The analysis of the literature review [11] allowed us to compare the RUP methodology against other software development methodologies, in order to use the best option

for the present work. From here, the RUP methodology could be identified as the best alternative due to its adaptability. On the other hand, in the investigation carried out, a series of findings were found due to the survey carried out where 79 percent were satisfied, coinciding with the author [30], regarding the use of a tool for detecting respiratory diseases in mobile phones, notwithstanding the research carried out by the author [14], it is very different since they used an interview looking at the opinion of the users involved. Regarding the prototype, the author [31], coupled the topic of a prototype on a web page and compare it with the research carried out, not having a coincidence since it limits us to only web prototypes without the use of an expert system. According to the efficiency results obtained, it agrees with the author [16], in that there isa fast response when displaying the results. Likewise, there is concindence with the function of diagnosing respiratory diseases with mobile application of the authors [15], however, they differ in the way of obtaining data on the symptoms presented by users.

VII. Conclusions and Future Works

In conclusion, in the present research work it was possible to successfully design a mobile health application for the registration and diagnosis of patients with respiratory diseases, by using the RUP methodology for its development, the quality of the product is guaranteed. In addition, it is supported by the survey of 51 users; which validated that the application meets the criteria of functionality, efficiency. efficacy and satisfaction. Likewise, the application allows people who suffer from respiratory diseases to register their information and have diagnosis from the comfort of your home. Regarding the limitations of the research, there is the lack of validation of the prototype by experts, since the collection of information about respiratory diseases was collected only from the bibliographic review. In addition, the data used for the survey has been limited to symptoms that occur in a small group of respiratory diseases; this is due to the time in which the data collection was carried out.

Based on the designed prototype, it is recommended to expand the functionalities of the application in future research in order to more accurately detect respiratory diseases. Future research is needed to expand the collective understanding of the application of other emerging technologies in health applications and to optimize diagnostic results in patients with respiratory diseases.

References

[1] A. D. Rio-Chillcce, L. Jara-Monge, and L. Andrade-Arenas, "Analysis of the use of videoconferencing in the learning process during the pandemic at a university in lima," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.01205102

[2] G. V. Torres, L. L. Aponte, and L. Andrade-Arenas, "Implementation of an expert system for automated symptom consultation in peru," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0121298

[3] F. Andrade-Chaico and L. Andrade-Arenas, "Projections on insecurity, unemployment and poverty and their consequences in lima's district san juan de lurigancho in the next 10 years," in *2019 IEEE Sciences and Humanities International Research Conference (SHIRCON)*, 2019, doi:10.1109/SHIRCON48091.2019.9024877, pp. 1–4.

[4] S. Caini, W. J. Alonso, A. Balmaseda, A. Bruno, P. Bustos, L. Castillo, C. D. Lozano, D. D. Mora, R. A. Fasce, W. A. F. D. Almeida, G. F. Kusznierz, J. Lara, M. L. Matute, B. Moreno, C. M. P. Henriques, J. M. Rudi, C. E. G. Séblain, F. Schellevis, J. Paget, W. Andrade, M. A. Becerra, M. Mejia, and A. W. Clara, "Characteristics of seasonal influenza a and b in latin america: Influenza surveillance data from ten countries," *PLoS ONE*, vol. 12, 2017, doi: 10.1371/journal.pone.0174592.

[5] L. Andrade-Arenas and C. Sotomayor-Beltran, "Evolution of acute respiratory infections in peru: A spatial study between 2011 and 2016." IEEE, 8 2019, doi = 10.1109/SCLA.2019.8905563, pp. 1–4.

[6] F. Mohd and N. I. E. Mustafah, ""hello, dr": A heatlhcare mobile application," 9 2021, doi:10.1109/ISAMSR53229.2021.9567764, pp. 20–23.

[7] J. Flores-Rodriguez and M. Cabanillas-Carbonell, "Mobile application for registration and diagnosis of respiratory diseases: a review of the scientific literature between 2010 and 2020." IEEE, 10 2020, doi:10.1109/EHB50910.2020.9280282, pp. 1–4.

[8] F. M. Rostami, B. N. Esfahani, A. M. Ahadi, and S. Shalibeik, "A review of novel coronavirus, severe acute respiratory syndrome coronavirus 2 (sars-cov-2)," *Iranian Journal of Medical Microbiology*, vol. 14, 2020, doi:10.30699/ijmm.14.2.154.

[9] D. Yang, T. Xu, X. Wang, D. Chen, Z. Zhang, L. Zhang, J. Liu, K. Xiao, L. Bai, Y. Zhang, L. Zhao, L. Tong, C. Wu, Y. Wang, C. Dong, M. Ye, Y. Xu, Z. Song, H. Chen, J. Li, J. Wang, F. Tan, H. Yu, J. Zhou, J. Yu, C. Du, H. Zhao, Y. Shang, L. Huang, J. Zhao, Y. Jin, C. A. Powell, Y. Song, and C. Bai, "A large-scale clinical validation study using ncapp cloud plus terminal by frontline doctors for the rapid diagnosis of covid-19 and covid-19 pneumonia in china," *medRxiv*, 2020.

[10] M. de Salud MINSA. (2022) Covid 19 en el perú - ministerio del salud. [Online]. Available: https://covid19.minsa.gob.pe/sala situacional.asp

[11] M. Sudarma, S. Ariyani, and P. A. Wicaksana, "Implementation of the rational unified process (rup) model in design planning of sales order management system," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 5, 2021, doi:10.29407/intensif.v5i2.15543.

[12] M. Berquedich, A. Berquedich, O. Kamach, M. Masmoudi, A. Chebbak, and L. Deshayes, "Developing a mobile covid-19 prototype management application integrated with an electronic health record for effective management in hospitals," *IEEE Engineering Management Review*, vol. 48, 2020.

[13] J. Flores-Rodriguez and M. Cabanillas-Carbonell, "Mobile application for registration and diagnosis of respiratory diseases: A review of the scientific literature between 2010 and 2020," in *2020 International Conference on e-Health and Bioengineering (EHB)*. IEEE, 2020, pp. 1–4.

[14] V. H. Arias Caballero, "Sistema experto para el diagnóstico de enfermedades respiratorias crónicas en el distrito la esperanza–provincia de trujillo," 2019.

[15] U. Sait, S. Shivakumar, G. L. K.V., T. Kumar, V. D. Ravishankar, and K. Bhalla, "A mobile application for early diagnosis of pneumonia in the rural context." IEEE, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/9033048/

[16] E. Candela-Mendoza, L. Cruz-Ipanaque, and J. Armas-Aguirre, "Mobile technology model to ambulatory healthcare information online using cloud platform," 2018, doi: 10.1109/INTERCON.2018.8526466.

[17] R. G. Figueroa, C. J. Solís, and A. A. Cabrera, "Metodologías tradicionales vs. metodologías Ágiles," *Universidad Técnica Particular de Loja, Escuela de Ciencias en Computación*, 2018.

[18] T. K. Tia, "Simulation model for rational unified process (rup) software development life cycle," *SISTEMASI*, vol. 8, 2019.doi:10.32520/stmsi.v8i1.420.

[19] J. D. L. Castillo, *Desarrollo de aplicaciones Android con Android Studio: Conoce Android Studio*. José Dimas Luján Castillo, 2019.

[20] J. F. Ramírez Rivas, "Implementación de lineamientos base de seguridad en bases de datos oracle y sql server en una entidad bancaria," 2019.

[21] T. Tia, I. Nuryasin, and M. Maskur, "Model simulasi rational unified process (rup) pada pegembagan perangkat lunak," *Jurnal Repositor*, vol. 2, 2020, doi:10.22219/repositor.v2i4.390.

[22] A. Rocha, A. I. de Sistemas e Tecnologias de Informacao, M. I. Systems, I. of Electrical, E. E. P. Section, I. of Electrical, and E. Engineers, *2021 16th Iberian Conference on Information Systems and Technologies (CISTI) : proceedings of CISTI'2021 - 16th Iberian Conference on Information Systems and Technologies : 23 to 26 of June 2021, Chaves, Portugal*.

[23] J. L. Leal, J. P. Rodríguez, and O. A. Gallardo, "Project time: Time management method for software development projects-analytical

summary," *Journal of Physics: Conference Series*, vol. 1126, p. 012030, 11 2018, doi: 10.1088/1742-6596/1126/1/012030.

[24] L. Chen and J. Y. Song, "Development of ahmes (automatical higher mathematics examination system) using rational unified process," *Mathematical Problems in Engineering*, vol. 2021, doi = 10.1155/2021/7952816, 2021.

[25] R. L. Ayala, N. V. Rosas, and L. Andrade-Arenas, "Implementation of a web system to detect anemia in children of peru," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, 2021, doi:10.14569/IJACSA.2021.0121299.

[26] A. Nasution, B. Efendi, and I. K. Siregar, "Pelatihan membuat aplikasi android dengan android studio pada smp negeri 1 tinggi raja," *Jurdimas (Jurnal Pengabdian Kepada Masyarakat) Royal*, vol. 2, 2019, 10.33330/jurdimas.v2i1.321.

[27] Y. Jiang, "Research on application value of computer software development in java programming language," vol. 1648, 2020, 10.1088/1742-6596/1648/3/032152.

[28] H. Choi, S. Lee, and D. Jeong, "Forensic recovery of sql server database: Practical approach," *IEEE Access*, vol. 9, 2021, 10.1109/ACCESS.2021.3052505.

[29] H. Pandowo, H. Tohari, and V. Amir, "Design and build the booking queue application android based soosoo resto," in *Journal of Physics: Conference Series*, vol. 1845, no. 1. IOP Publishing, 2021, p. 012012.

[30] A. Anand, D. Chamberlain, R. Kodgule, and R. R. Fletcher, "Pulmonary screener: A mobile phone screening tool for pulmonary and respiratory disease," in *2018 IEEE Global Humanitarian Technology Conference (GHTC)*, 2018, doi:10.1109/GHTC.2018.8601821, pp. 1–7.

[31] D. E. Liberato Bernal and R. M. Quilcat Peantes, "Sistema informático móvil inteligente para la detección temprana y control de enfermedades respiratorias en pacientes del sector privado de salud en la ciudad de trujillo," 2021.

Summarizing Event Sequence Database into Compact Big Sequence

Mosab Hassaan
Faculty of Science, Benha University, Egypt

***Abstract*—Detecting the core structure of a database is one of the most objective of data mining. Many methods do so, in pattern set mining, by mining a small set of patterns that together summarize the dataset in efficient way. The better of these patterns, the more effective summarization of the database. Most of these methods are based on the Minimum Description Length principle. Here, we focus on the event sequence database. In this paper, rather than mining a small set of significant patterns, we propose a novel method to summarize the event sequence dataset by constructing compact big sequence namely, BigSeq. BigSeq conserves all characteristics of the original event sequences. It is constructed in efficient way via the longest common subsequence and the novel definition of the compatible event set. The experimental results show that BigSeq method outperforms the state-of-the-art methods such as Gokrimp with respect to compression ratio, total response time, and number of detected patterns.**

***Keywords*—Sequence data; compressing patterns mining; minimum description length**

I. Introduction

Detecting the key patterns from a database is one of the main objectives of data mining. There are many studies for mining all patterns that satisfy some constraints (such patterns may be frequent patterns as in PrefixSpan [11], CM-SPADE [19], PRISM [15], and [20] , or closed patterns as in [24][7][14][2], or maximal patterns as in [3], [23]). Rather than mining all patterns, existing methods mining a set of patterns that is significat for summarizing the dataset. There are many methods to define this significant patterns. One of these methods is the Minimum Description Length (MDL) principle [21][12][6][22] which has proven to be particularly the winner one.It is based on the insight that any regularity in the dataset can be used to compress the dataset. Note that, the more we can compress, the more regularity we have found. More details about MDL are described in next section.

For itemsets data, Krimp [13] is based on MDL principle. For sequence data, the authors of SeqKrimp [8][9], Gokrimp [8][9], and SQS [18] used MDL principle to compress the sequence data. More details about these algorithms are illustrated in the related work section (Section III).

In this paper, we focus on the event sequence data. Our objective is to search for a summary of the given event data sequences. The size of this summary must be very small compared to the size of the event sequence dataset. Also this summary must converse the all characteristics of the original event sequences. The existing methods mine a significant patterns that compress the dataset well. Some of these methods generate the sequential patterns as a first phase. Then the significant patterns are selected with respect to MDL as a second phase. Note that the significant patterns is only a small subset of the set of all the sequential patterns and the process of mining all sequential patterns is very expensive process. Therefore, the other existing methods devise some effective pruning methods to prune the ineffective parts of the search space that do not contain any significant pattern. Unfortunately, the process of the pruning the ineffective parts of the search space consumes more time if it not used efficient techniques.

Contribution. From above, all existing methods apply the mining process to search for the significant patterns. In contrast, our proposed method donot apply the mining process. Instead of, all event sequences in the dataset are merged into only one compact big sequence. In other words, our proposed method detects only one significant pattern which is the compact big sequence. Note that, the detected big sequence must be compact as much as possible. Therefore, we introduce an efficient method for constructing the big sequence to reduce the size of big sequence as much as possible. The construction method is based on the longest common subsequence and the novel definition of the compatible event set. Our compact big sequence converses the all characteristics of the original event sequences via preserving the order of events as in the dataset and also associating with each event in the big sequence a list of sequence ids that contains this event. To confine the larger size of the lists of sequence ids, we can represent them as sets of bit-vectors. Here, the consecutive zeros in the sets of bits-vector are compressed in efficient way.

Organization. This paper is organized as follows. Section II defines the preliminary concepts. Section III presents the related work. Section IV presents our proposed algorithm. Section VI reports the experimental results. Finally, Section VII concludes the paper.

II. Preliminary Concepts

Let $E = \{e_1, e_2, ..., e_m\}$ be a set of m distinct events. Event sequence $S =< u_1, u_2, ..., u_l >$ over E is ordered list such that $u_i \in E$. Event sequence $W = \{w_1, w_2, ..., w_h\}$ is subsequence of the event sequence S if there are h integers $(j_1, j_2, ..., j_h)$ such that $1 \leq j_1 < j_2 < ... < j_h \leq l$ and $w_1 = s_{j_1}$, $w_2 = s_{j_2}$,... ,$w_h = s_{j_h}$. Event sequence with length l is called an l-sequence. Event sequence database $\mathcal{D}$ = $\{S_1, S_2, ..., S_n\}$ is a set of event sequences where $|\mathcal{D}|$ = n. For example, consider Table I which contains an example of event sequence database $\mathcal{D}$ with $|\mathcal{D}|$ = 8. The sequence S_5 = $ABCB$ is subsequence of the sequence S_1 = $ABCBC$ ($S_5 \sqsubseteq S_1$). Also we can said S_1 is supersequence of S_5.

Definition 2.1: **Longest Common Subsequence.** Given two event sequences X and Y, the longest common

TABLE I. EVENT SEQUENCE DATABASE, $\mathcal{D}$

Sid	Sequence
S_1	$ABCBC$
S_2	$ABAA$
S_3	$CABAC$
S_4	CAC
S_5	$ABCB$
S_6	$CBAC$
S_7	$BCAB$
S_8	$ACBBA$

subsequence between X and Y denoted as $lcs(X, Y)$ is a longest sequence Z that is a subsequence of both X and Y.

Problem Definition: Given event sequence database $\mathcal{D}$, the objective is to find a summary, $\mathcal{S}$ of $\mathcal{D}$ such that $\mathcal{S}$ conserves all characteristics of $\mathcal{D}$ and the size of $\mathcal{S}$ is sharply less than the size of $\mathcal{D}$. ($|\mathcal{S}| << |\mathcal{D}|$).

III. RELATED WORK

In the beginning, we discuss the minimum description length in details as follows.

The Minimum Description Length

The minimum description length (MDL) principle [21][12][6][22] widely used in text compression. It used as a method for selecting a set of compressive patterns. If these patterns are used as a dictionary then we have {the potential} to maximally compress the dataset into a compact pattern encoding. In other words, these patterns represent the dataset in efficient way. Unfortunately, the process of selecting such patterns that based on MDL is NP-hard problem. Given a set of models $\mathcal{M}$, the MDL principle states that the winner model $M \in \mathcal{M}$ for the dataset $\mathcal{D}$ is the best model that provides the lossless compression. Formally, we optimize $Len(\mathcal{D}, M) = Len(M) + Len(\mathcal{D} \setminus M)$ where $Len(M)$ is the length in bits of the description of $\mathcal{M}$ and $Len(\mathcal{D} \setminus M)$ is the length in bits of the dataset when compressed with model M. MDL was applied to detect compressed frequent patterns from itemsets and sequences data. In next sections, we discuss the algorithms that based on MDL.

For itemsets data, there is algorithm called Krimp [13] that based on MDL principle. This algorithm is effective in solving the redundancy issue in the descriptive pattern mining. For sequence data, the authors of SeqKrimp [8][9] used MDL principle to compress the sequence data. This algorithm contains two steps. The first step generates the sequential patterns as candidates by using existing sequence mining method. The second step greedily checks the candidate set to find the useful patterns which together minimizes the description length. The SeqKrimp algorithm has two main disadvantages which are the process of generating the candidates is expensive and the patterns that do not belong to candidate set have no chance to be selected even if they have ability to minimizes the description length.

The authors of Gokrimp [8][9] mine a set of non-redundant sequential patterns that compress the sequence data using the MDL principle. GoKrimp do not generate candidates as in SeqKrimp. Instead of, it directly mines compressed useful patterns by greedily extending a pattern until no additional compression benefit added. To taming the hardness of the checks for additional compression benefit of an extension, Gokrimp proposed a dependency test which only selects related events for extending a given pattern.

As in GoKrimp, SQS [18] also directly mines the compressed patterns from the sequence dataset. The patterns are constructed iteratively. In each iteration, the pattern is selected if it achieves the largest MDL gain among the possible patterns. Note that, each iteration requires at least one scan of the sequence dataset.

IV. PROPOSED ALGORITHM

The method is based on the observation that the most event sequences in real dataset share the same subsequences. To avoid the overhead of duplicated computations, we propose big sequence method that merges all event sequences in the dataset into one big sequence abbreviated as BigSeq. The construction method of BigSeq is one of main operations in our algorithm. BigSeq must be compact and efficient. At the same time, it must conserve all characteristics of the original dataset.

To construct compact BigSeq, we should propose an efficient method to reduce the size of BigSeq as much as possible. Thus, we will propose a new efficient method to construct BigSeq. Next we discuss the steps of the construction method on the sequence dataset of Table I.

First, we select any sequence S in the sequence database, $\mathcal{D}$ (see Table I) as initial value of BigSeq. Suppose we selected the first sequence $S_1 \in \mathcal{D}$. Then the BigSeq is $ABCBC$. As we will see, some events will be inserted into the current BigSeq to generate the final BigSeq. Therefore, we set a temporary index for each event in the current BigSeq as follows. The temporary indices of events in the current BigSeq will be $i_1 i_2 i_3 i_4 i_5$ with $i_1 << i_2 << i_3 << i_4 << i_5$. We can assume the following $i_j = i_{j-1} + (j-1).\epsilon$ with $2 \leq j \leq 5$ and $\epsilon \geq 1$. For example, $i_3 = i_2 + 2\epsilon$ After generating the final BigSeq, we will set the actual value for each temporary index, i_j.

Second, for each remaining sequence S' in $\mathcal{D}$, compute the longest common subsequence between S' and BigSeq, namely $LCS(S', BigSeq)$. After that we store the positions in BigSeq for each event that belong to LCS and store also the remaining events in S', that do not belong to $LCS(S', BigSeq)$ (Note that these remaining events will be further inserted in BigSeq). For example, let S' be the sixth sequence, $S' = S_6 = CBAC$. We have $LCS(S', BigSeq) = LCS(CBAC, ABCBC) = CBC$. The positions of the three events (C, B, and C) of $LCS(S', BigSeq)$ in BigSeq are i_3, i_4, and i_5, respectively. We store these positions. Also, we store the remaining event, A, in S' that does not belong to $LCS(S', BigSeq)$. This remaining event, A, will be further inserted in BigSeq. The remaining events of each remaining sequence must be inserted in the correct position in the BigSeq. Therefore, we will associate with each remaining event e_r a range of positions in BigSeq. We expect that e_r will fall within this range in BigSeq. We call this range an **E**xpected **R**ange of **P**ositions for event e_r, namely $ERP(e_r)$. Recall let $S' = S_6 = CBAC$ then we have only one remaining event A. The position of the event A in S' falls between the positions of two events

B and C. Note that these two events (B and C) belong to $LCS(S', BigSeq)$ and their positions in BigSeq are i_4 and i_5. Therefore, we have $ERP(A) =]i_4, i_5[$. As a consequence, we should insert the remaining event A in BigSeq at a new position between i_4 and i_5. Table II shows the expected range of positions for each remaining event e_r, $ERP(e_r)$.

Finally, indeed, we do not insert each remaining event in BigSeq instead of we cluster the remaining events into compatible event sets. After that we insert only one **rep**resntative **e**vent, e_{rep}, for each compatible event set into BigSeq at a specific position p. This position p must belong to the expected range of positions of every event in the compatible event set of e_{rep}. See next definition of compatible event set and see next example.

Definition 4.1: **Compatiable Event Set.**
The event set is called compatible event set if the events in this set satisfy the next three conditions:

1) They have the same label;
2) They donot belong to the same event sequence;
3) The insertion of their expected range of positions in BigSeq is not empty.

Example 4.1: Given the event sequence database in Table I. Table II reports the initial value of BigSeq (S_1 [The first row]), the remaining sequences (S_2, S_3, S_4, S_5, S_6, S_7, S_8 [The first column]), the events of each remaining sequence S' that belong to $LCS(S', BigSeq)$ [The second column], and $ERP(e_r)$ for any remaining event, e_r ($e_r \notin LCS(S', BigSeq)$) [The third column].

Note that the underlined events in the first and second columns belong to $LCS(S', BigSeq)$ and the parameter $\delta \geq 1$. To distinguish among the remaining events (e_r in the third column of Table II) that have the same label, we assign superscripts for these events as follows. A^{km} means the m-th remaining event in the sequence k.

Now we will determine the compatible sets of remaining events. The remaining event can be belonged to more than one compatible event set. In this case, we add this remaining event to only one compatible event set. From the definition of compatible event set, If two or more different remaining events belong to the same event sequence then we must add them to different sets of compatible events. For example, since the two different remaining events, A^{21} and A^{22} belong to the same event sequence (the second event sequence, S_2), they must be added to two different sets of comaptiable events. Based on the definition of compatible event set and $ERP(e_r)$ in Table II, we have three **co**mpatible sets of **r**emaining **e**vents, $core = \{core_1, core_2, core_3\}$, where $core_1 = \{A^{21}, A^{31}, A^{41}, A^{71}\}$, $core_2 = \{A^{22}, A^{32}, A^{61}, A^{82}\}$, and $core_3 = \{C^{81}\}$.

Next we will discuss the computations of these three compatible sets of remaining events in details and how we conserve all characteristics of the original event sequence in the final BigSeq with respect to the event sequence database and $ERP(e_r)$ in Tables I and II, respectively.

The first compatible set of remaining events is $core_1 = \{A^{21}, A^{31}, A^{41}, A^{71}\}$. Note that all events in $core_1$ have the same label, A (Condition 1 in Definition 4.1) and they do not belong to the same sequence but they belong to sequences S_2, S_3, S_4, and S_7 respectively (Condition 2 in Definition 4.1). We have $ERP(A^{21}) = [p, p + \delta[$, where $p > i_2$. Recall, because $i_3 > i_2$, we can set $p = i_3$. Now $ERP(A^{21}) = [i_3, i_3 + \delta[$. Recall, $\delta \geq 1$ then we can set δ = 2. In other words, $ERP(A^{21}) = [i_3, i_3 + 2[$. The other expected range of positions are $ERP(A^{31}) =]i_3, i_4[$, $ERP(A^{41}) =]i_3, i_5[$, and $ERP(A^{71}) =]i_3, i_4[$. As the result, we have $ERP(A^{21}) \cap ERP(A^{31}) \cap ERP(A^{41}) \cap ERP(A^{71}) \neq \phi$ (Condition 3 in Definition 4.1). Thus, we insert in BigSeq at position $i_3 + 1$ only one event with label A (e_{rep1}) as representative for $core_1$. Note that we select the position $i_3 + 1 \neq i_4$, since $i_3 + 1 \in ERP(A^{21}), ERP(A^{31}), ERP(A^{41}), and\ ERP(A^{71})$.

The second compatible set of remaining events is $core_2 = \{A^{22}, A^{32}, A^{61}, A^{82}\}$. Note that $core_2$ satisfy coditions 1 and 2 in Definition 4.1 since all events in $core_2$ have the same label, A. the events in $core_2$ do not belong to the same sequence but they belong to sequences S_2, S_3, S_6, and S_8 respectively (Condition 2 in Definition 4.1). We have $ERP(A^{22}) = [p + \delta, \infty[= [i_3 + 2, \infty[$. The other expected range of positions are $ERP(A^{32}) =]i_4, i_5[$, $ERP(A^{61}) =]i_4, i_5[$, and $ERP(A^{82}) = [i_4, \infty[$. As the result, we have $ERP(A^{22}) \cap ERP(A^{32}) \cap ERP(A^{61}) \cap ERP(A^{82}) \neq \phi$ (Condition 3 in Definition 4.1). We insert in BigSeq at position $i_4 + 1$ only one event with label A (e_{rep2}) as representative for $core_2$. Note that we select the position $i_4 + 1 \neq i_5$, since $i_4 + 1 \in ERP(A^{22}), ERP(A^{32}), ERP(A^{61}), and\ ERP(A^{82})$.

Finally, the third compatible set of remaining events is $core_3 = \{C^{81}\}$ with $ERP(C^{81}) =]i_1, i_2[$. Since $core_3$ has only one event then it is compatible set that satisfy the three conditions in Definition 4.1. Thus, we insert the event C in BigSeq as representative for $core_3$ [e_{rep3}] at position between i_1 and i_2. In other words, we can insert C in BigSeq at position $i_1 + 1 \neq i_2$ such that $i_1 + 1 \in]i_1, i_2[$.

The initial BigSeq with its indices and the final BigSeq with its indices are reported at Table III(a) and Table III(b), respectively. The final BigSeq is $ACBCABAC$.Note that we insert the three representative C, A, and A in BigSeq at position $i_1 + 1$ (between i_1 and i_2), $i_3 + 1$ (between i_3 and i_4), and $i_4 + 1$ (between i_4 and i_5). Here the size of the final BigSeq is 8 after inserting the three representatives. Therefore, the actual indices of the events in the final BigSeq will be from 1 to 8 (i.e. 1, 2, 3, 4, 5, 6, 7, and 8). See the next definition for the size of the final BigSeq.

Definition 4.2: **The Final BigSeq Size.**
The Final BigSeq Size is $|final\_BigSeq|$ = $|initial\_BigSeq|$ + $|core|$, where $|core|$ is the count of compatible sets of remaining events, where $initial\_BigSeq$ is the initial value of BigSeq.

For example, with respect to the event sequence database and the data in Tables I and II respectively, we have the following $|final\_BigSeq|$ = $|initial\_BigSeq|$ + $|core|$ = $|S_1|$ + $|core|$ = 5 + 3 = 8.

From definition 4.2, to reduce the final BigSeq Size, we should reduce the count of compatible sets of the remaining events as much as possible.

To conserve all characteristics of the original event sequence in the final BigSeq, we should associate with each event e in BigSeq a list of sequence ids that contains the

TABLE II. EXPECTED RANGE OF POSITIONS FOR EACH REMAINING EVENT e_r, $ERP(e_r)$

Initial Value of BigSeq = S_1 = $ABCBC$ (the first sequence in $\mathcal{D}$) and its indices are $i_1i_2i_3i_4i_5$		
S': Remaining Seq	**$e \in LCS(S', BigSeq)$ with pos. in BigSeq**	$ERP(e_r)$
S_2: $\underline{A}\underline{B}AA$	$\underline{A}$ and $\underline{B}$ with pos. i_1 and i_2	$ERP(A^{21}) = [p, p+\delta[$, $p > i_2$ - $ERP(A^{22}) = [p+\delta, \infty[$
S_3: $\underline{C}A\underline{B}A\underline{C}$	$\underline{C}$, $\underline{B}$, and $\underline{C}$ with pos. i_3, i_4, and i_5	$ERP(A^{31}) =]i_3, i_4[$ and $ERP(A^{32}) =]i_4, i_5[$
S_4: $\underline{C}A\underline{C}$	$\underline{C}$ and $\underline{C}$ with pos. i_3 and i_5	$ERP(A^{41}) =]i_3, i_5[$
S_5: $\underline{ABCB}$	$\underline{A}$, $\underline{B}$, $\underline{C}$, and $\underline{B}$ with pos. i_1, i_2, i_3, and i_4	NULL
S_6: $\underline{C}\underline{B}A\underline{C}$	$\underline{C}$, $\underline{B}$, and $\underline{C}$ with pos. i_3, i_4, and i_5	$ERP(A^{61}) =]i_4, i_5[$
S_7: $\underline{BC}A\underline{B}$	$\underline{B}$, $\underline{C}$, and $\underline{B}$ with pos. i_2, i_3, and i_4	$ERP(A^{71}) =]i_3, i_4[$
S_8: $A\underline{C}\underline{B}\underline{B}A$	$\underline{A}$, $\underline{B}$, and $\underline{B}$ with pos. i_1, i_2, and i_4	$ERP(C^{81}) =]i_1, i_2[$ and $ERP(A^{82}) =]i_4, \infty[$

TABLE III. BIGSEQ CONSTRUCTION

Temp. Pos.	i_1	i_2	i_3	i_4	i_5
BigSeq Events	A	B	C	B	C

(a) Initial BigSeq

Temp. Pos.	i_1	i_1+1	i_2	i_3	i_3+1	i_4	i_4+1	i_5
Actual Pos.	1	**2**	3	4	**5**	6	**7**	8
BigSeq Events	A	**C**	B	C	**A**	B	**A**	C

(b) Insertion of the Three Representatives of the Three Compatiable Sets in BigSeq

TABLE IV. STEPS OF BIGSEQ CONSTRUCTION WITH SEQUENCE ID LIST

Pos.	i_1	i_2	i_3	i_4	i_5
BigSeq Events	A	B	C	B	C
Seq. Id List	1	1	1	1	1

(a) Initial BigSeq with Id List of the First Sequence

Pos.	i_1	i_2	i_3	i_4	i_5
BigSeq Events	A	B	C	B	C
Seq. Id List	1	1	1	1	1
	2	2	3	3	3
	5	5	4	5	4
	8	7	5	6	6
		8	6	7	
			7	8	

(b) Addition of Id List for each Event $e \in GCD(S', BigSeq)$

Temp. Pos.	i_1	$\mathbf{i_1+1}$	i_2	i_3	$\mathbf{i_3+1}$	i_4	$\mathbf{i_4+1}$	i_5
Actual Pos.	1	**2**	3	4	**5**	6	**7**	8
BigSeq Events	A	**C**	B	C	**A**	B	**A**	C
Seq. Id List	1	8	1	1	2	1	2	1
	2		2	3	3	3	3	3
	5		5	4	4	5	6	4
	8		7	5	7	6	8	6
			8	6		7		
				7		8		

(c) Insertion of the Three Representatives for the Three Compatible Sets in BigSeq with their Id List

event e, namely $e.id\_list$ as follows. First, since we select the first sequence as the initial value for BigSeq, we will add 1 (the id of the first sequence) to $e.id\_list$ for each event $e \in initial\_BigSeq = S_1$ [see Table IV (a)]. Second, suppose the case that the event $e \in LCS(S', BigSeq)$, where S' is a remaining sequence (i.e. $e \in S'$ and $e \in BigSeq$). In this case, we add the id of S' to $e.id\_list$ for each event $e \in BigSeq$ [see Table IV (b)]. Finally, we have three representative events for the three compatible sets of remaining events. As we mentioned before, we inserted the three representatives, A, A, and C in BigSeq at positions i_3+1, i_4+1, and i_1+1 respectively to generate the final BigSeq. For each representative event, e_{rep}, for the compatible set of remaining events, $core_k$, we add to $e_{rep}.id\_list$ the id of the event sequence that contains the remaining event e_r, for every $e_r \in core_k$ with k = 1, 2, and, 3 [see Table IV (c)].

Next algorithm outlines the BigSeq construction with sequence Id List.

Algorithm: BigSeq Construction with Sequence Id List

Input: Event sequence database, $\mathcal{D}$
Output: $BigSeq$ with $e.id\_list$ for each event $e \in BigSeq$.

1. Select an event sequence $S \in \mathcal{D}$ as $BigSeq$
   // Initial value of BigSeq = S
2. Add the id of S to $e.id\_list$ for each $e \in BigSeq$
3. $\mathcal{D} = \mathcal{D} - S$
4. $ERP = \{\}$
   //the set of expected range of positions
5. **for** each event sequence $S' \in \mathcal{D}$ **do**
6. $lcs = LCS(S', BigSeq)$
7. **for** each event $e \in lcs$ **do** // $e \in S'$ and $e \in BigSeq$
8. Add the id of the sequence S' that contains the event e to $e.id\_list$ in $BigSeq$
9. **end for**
10. **for** each event $e_r \in S'$ and $e_r \notin lcs$ **do**
    //e_r is remaining event
11. Compute $ERP(e_r)$
12. $ERP = ERP \cup ERP(e_r)$
13. **end for**
14. **end for**
15. Find the compatible sets of remaining events, $core$, based on Definition 4.1 and ERP
16. **for** each compatible set $core_k \in core$
17. Insert the representative event, e_{rep}, for $core_k$ into $BigSeq$ at position $p \in ERP(e') \; \forall \; e' \in core_k$
18. Add to $e_{rep}.id\_list$ the id of the event sequence that contains the remaining event $e_r \; \forall \; e_r \in core_k$
19. **end for**
20. **return** $BigSeq$ // The final BigSeq with $e.id\_list$

TABLE V. BIGSEQ WITH BIT-VECTORS

Pos.	1	2	3	4	5	6	7	8
$e \in BigSeq$	A	C	B	C	A	B	A	C
$B(e) = \{B_1\}$	{10010011}	{10000000}	{11010011}	{01111101}	{01001110}	{11110101}	{10100110}	{00101101}

for each event $e \in BigSeq$

V. COMPRESSING EVENTS SEQUENCES DATABASE USING $BigSeq$ METHOD

The objective of this paper is to compress the event sequence database in efficient way such that we conserve all characteristics of the original database. In other words, we will compress the event sequence database into compact BigSeq with sequence id_lists. But when the size of event sequence database is large, the id_list size of each event in the corresponding BigSeq will be large. To confine the larger size of these id_lists, we can represent $e.id\_list$ of each event $e \in BigSeq$ as a set of bit-vectors, $B(e) = \{B_1, B_2, ..., B_m\}$, where each B_i is 8 length bit-vector (i.e. each B_i occupy 1 byte in memory) and suppose that the maximum $id \in e.id\_list$ is n then $m = |B(e)| = n/8$. Each position in each B_i corresponds to event sequence $S_{id} \in \mathcal{D}$ where $id \in [8 \times (i-1) + 1, 8 \times i]$. The bit at position j in B_i represents the presence or absence of the event $e \in BigSeq$ in the event sequence $S_{8\times(i-1)+j}$. See next example.

Example 5.1: The first event in BigSeq (in Table V) is $e_1 = A$ with the set of bit-vectors $B(e_1) = \{B_1\} = \{10010011\}$. Note that $B(e_1)$ contains only one bit-vector, B_1, since the maximum id in $e_1.id\_list$ = 8 (i.e. $n = 8$), thus $m = n/8 = 1 = |B(e)|$. The bits in B_1 represent the presence or absence of the event e_1 in the event sequences that have $id \in [1, 8]$. The bit at position 1 in B_1 is one, this means that $e_1 \in S_1 = S_{8\times(1-1)+1}$, etc. Given the final BigSeq with id_lists in Table IV(c), its corresponding BigSeq with bit-vectors is reported in Table V.

A. Compression Benefit

Suppose each event e occupy 1 byte in memory, then the size (in terms of bytes) of the original event sequence database, $\mathcal{D}$ (in Table I) is 34 bytes ($\mathcal{D}$ contains 34 events). Recall each B_i occupy also 1 byte in memory, therefore the size (in terms of bytes) of the BigSeq with bit-vectors (in Table V) is $|BigSeq|$ + $|B(e)|$ = 8 + 8 = 16 bytes. We can use the compression ratio to measure how well the data is compressed. The compression ratio calculated by dividing the data size before compression with the size after compression. In the above example, the compression ratio is $34/16$ = 2.125. In other words, the space saving (%) is (1 - (compressed size / uncompressed size)) $\times$ 100 = $1 - (16/34) \times 100$ = 52.9%. Here, we have an optimization that based on the observation that there are many consecutive zeros in each row of the bit-vectors. This is clearly grossly inefficient. Therefore, we can compress these consecutive zeros in efficient way as follows. Given array of bits (0 and 1), Bit_Arr, and paramter n. The output is the same as the input except for consecutive zeros. Note that, may be there are many sets of consecutive zeros in Bit_Arr. For each set of consecutive zeros, CZ, we do the following. If $|CZ| \leq n + 2$, we do nothing. Otherwise, we compress CZ into compressed CZ with size $n+2$ bits. The first and the second bits in the compressed CZ are 0 (indicator for compressing CZ) and 1 (indicator for doing the compression) respectively. The other n bits in the compressed CZ indicate how many times of zeros were repeated consecutively.

In the experimental results section, we will show the better compression ratio of BigSeq method against the state-of-art algorithm, GoKrimp, on many real datasets.

VI. EXPERIMENTAL EVALUATION

This section reportes the results of experiments on many real dataset. We compare the performance of Our proposed method, namely BigSeq with GoKrimp algorithm [8] [9]. Here, we exclude the two algorithms SeqKrimp and SQS from this experiment since GoKrimp algorithm outperforms them by one to two orders of magnitude. BigSeq is implemented in standard C++ with STL library support and compiled with GNU GCC. Experiments were run on laptop with Intel i3 2.4 GHz and 8G memory running Linux.

A. Datasets

Experimental evaluation are performed on a group of real datasets as follows. We used five real datasets namely, msnbc [10], Gene [16], TCAS [17] [4], Activity [1], and JBoss [5]. The corresponding information of these real datasets is summarized in Table VI, where $|\mathcal{D}|$ represents the number of sequences, $|E|$ is the number of the events, min_L, max_L and avg_L denote the minimum length, maximum length and average length of the sequences respectively.

TABLE VI. SUMMARY STATISTICS OF THE REAL DATASETS USED IN THE EXPERIMENTS

Dataset	$\lvert\mathcal{D}\rvert$	$\lvert E\rvert$	min_L	max_L	avg_L
msnbc	31790	18	9	100	13.33
Gene	2942	5	41	216	86.53
TCAS	1578	75	8	70	36
Activity	35	10	12	43	21.14
JBoss	28	64	51	125	91

B. Effect of Optimization

In this experiment, we show the effect of optimization of compressing the consecutive zeros, namely Opt, with respect to the compression ratio. Table VII reports the compression ratio of BigSeq with and without Opt on the three datasets (Gene, TCAS, and JBoss). From this table, BigSeq with Opt has the better compression ratio in all datasets. Note that the larger compression ratio is the better compression we have.

TABLE VII. EFFECT OF OPTIMIZATION WITH RESPECT TO COMPRESSION RATIO

Dataset	BigSeq with Opt	BigSeq without Opt
Gene	2.428	1.348
TCAS	3.384	1.585
JBoss	3.303	2.700

C. Performance of BigSeq against GoKrimp

From the previous experiment, BigSeq with Opt has the best performance with respect to compression ratio. Therefore, in this experiment, we will use BigSeq with Opt and we will call it as BigSeq for abbreviation.

The proposed method, BigSeq is evaluated according to the following criteria:

- *Compression Ratio:* To measure how well the dataset is compressed using BigSeq.
- *Total Response Time:* To measure the efficiency of BigSeq.
- *The Number of Patterns:* The number of detected patterns that used for compression.

1) Compression Ratio: Table VIII reports the compression ratio of the two algorithms on the five datasets. Recall, the larger compression ratio is the better compression we have. The BigSeq algorithm shows a better compression ratio in all datasets. For example, in Gene dataset, the compression ratio of BigSeq is 2.428 while the compression ratio of GoKrimp is 1.251.

TABLE VIII. COMPRESSION RATIO OF THE TWO ALGORITHMS (BIGSEQ AND GOKRIMP)

Dataset	BigSeq	GoKrimp
msnbc	1.648	1.123
Gene	2.428	1.251
TCAS	3.384	2.951
Activity	1.520	1.077
JBoss	3.303	1.541

2) Total Response Time (Sec): Table IX reports total response time (Sec) of the two algorithms on the five datasets. The BigSeq algorithm has the best execution time on all datasets. On msnbc dataset, Gene dataset, TCAS dataset, Activity dataset, and JBoss dataset, BigSeq outperforms GoKrimp by more than two orders of magnitude, more than one order of magnitude, more than three orders of magnitude, approximately three factors, and more than one order of magnitude respectively.

3) Number of Patterns: Table X reports the number of patterns that used for compression by the two algorithms on the five datasets. Note that BigSeq used only one pattern for all datasets. This pattern is the compact BigSeq itself.

TABLE IX. TOTAL RESPONSE TIME (SEC) OF THE TWO ALGORITHMS (BIGSEQ AND GOKRIMP)

Dataset	BigSeq	GoKrimp
msnbc	1.32	313
Gene	1.1	45.2
TCAS	0.135	199
Activity	0.066	0.239
JBoss	0.072	2

TABLE X. THE NUMBER OF PATTERNS OF THE TWO ALGORITHMS (BIGSEQ AND GOKRIMP)

Dataset	BigSeq	GoKrimp
msnbc	1	27
Gene	1	6
TCAS	1	33
Activity	1	2
JBoss	1	5

VII. CONCLUSION

In this paper, we focus on summarizing the event sequence dataset. Existing methods summarize the event sequence dataset by mining a significant patterns that compress the dataset well. In contrast, the novel proposed method, BigSeq summarizes the event sequence dataset by merging the event sequences into compact big sequence. The construction of the compact big sequence is done via the longest common subsequence and the novel definition of the compatible event set. Our compact big sequence converses the all characteristics of the original event sequences. Experimental results show that BigSeq method can achieve better performance than the state-of-the-art methods such as GoKrimp in terms of compression ratio, total response time, and number of detected patterns. As future work, we plan to adapt the BigSeq method for mining the frequent, closed, and maximal patterns in the event sequence dataset.

ACKNOWLEDGMENT

I wish to express my deep gratitude to my mentor Prof Dr. Karam Gouda. I am very grateful to my parents, my wife, my brother, and my sisters for their continuous moral support and encouragement.

REFERENCES

[1] A. Asuncion and D. Newman. *UCI machine learning repository* University of California, Irvine, School of Information and Computer Sciences, 2007. [Online]. Available: http://archive.ics.uci.edu/ml

[2] B. Le, H. Duong, T. Truong, and P. Fournier-Viger. *FCloSM, FGenSM: Two New Algorithms for Efficiently Mining Frequent Closed and Generator Sequences using Local Pruning Strategy*. Knowledge and Information Systems (KAIS), Springer, 53(1):71-107, 2017.

[3] B. Vo, S. Pham , T. Le, Z-H Deng. *A novel approach for mining maximal frequent patterns*. Expert Syst Appl 73:178-186, 2017

[4] C. Liu, X. Yan, L. Fei, J. Han, and S. P. Midkiff. *SOBER: statistical model-based bug localization*. In ACM SIGSOFT ESEC-FSE, 2005.

[5] D. Lo, S.-C. Khoo, and Chao Liu. *Efficient Mining of Iterative Patterns for Software Specification Discovery*. In, KDD, 2007.

[6] F. Bariatti, P. Cellier, and S. Ferré. *GraphMDL: graph pattern selection based on minimum description length*. In: International symposium on intelligent data analysis (IDA). Springer, 54-66, 2020.

[7] F. Fumarola, P. F. Lanotte, M. Ceci, and D. Malerba. *CloFAST: closed sequential pattern mining using sparseand vertical id-lists*. Knowl Inf Syst, 48:429-463, 2016.

[8] H. T. Lam, F. Mörchen, D. Fradkin, and T. Calders, *Mining compressing sequential patterns*, In SDM, 2012.

[9] H. T. Lam, F. Mörchen, D. Fradkin, and T. Calders. *Mining compressing sequential patterns. Statistical Analysis and Data Mining*, 7(1):34-52, 2014.

[10] http://www.msnbc.com

[11] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, M.-C. Hsu. *Mining sequential patterns by pattern-growth: The prefixspan approach*. IEEE Trans. Knowl. Data Eng. 16(11) 1424-1440, 2004.

[12] J. Rissanen, *Modeling by shortest data description*, Automatica, 14(1):465-471, 1978.

[13] J. Vreeken, M. van Leeuwen, and A. Siebes. *KRIMP: Mining itemsets that compress.* Data Min. Knowl. Disc., 23(1):169-214, 2011.

[14] J. Wang and J. Han. *BIDE: efficient mining of frequent closed sequences* In ICDE, 2004.

[15] K. Gouda, M. Hassaan, and M. J. Zaki. *Prism: An effective approach for frequent sequence mining via prime-block encoding*. Journal of Computer and System Sciences 76:88-102, 2010.

[16] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou. *Improved and promising identification of human micrornas by incorporating a high-quality negative set* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11(1) 192-201, 2014.

[17] M. Hutchins, H. Foster, T. Goradia, and T. Ostrand. *Experiments of the effectiveness of dataflow- and controlflow-based test adequacy criteria*. In ICSE, 1994.

[18] N. Tatti and J. Vreeken. *The long and the short of it: Summarizing event sequences with serial episodes.* In KDD, 462-470. ACM, 2012.

[19] P. Fournier-Viger, A. Gomariz, M. Campos, and R. Thomas. *Fast vertical mining of sequential patterns using co-occurrence information*. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 40-52, 2014.

[20] P. Fournier-Viger, C.-W. Wu, A. Gomariz, and V. S. Tseng. *VMSP: Efficient Vertical Mining of Maximal Sequential Patterns*. Proc. 27th Canadian Conference on Artificial Intelligence (AI), Springer, LNAI, pp. 83-94, 2014.

[21] P. Grünwald, *The Minimum Description Length Principle*, MIT Press, 2007.

[22] T. Makhalova, S. O. Kuznetsov, and A. Napoli *Mint: MDL-based approach for Mining INTeresting Numerical Pattern Sets*. Data Mining and Knowledge Discovery 36:108-145, 2022.

[23] Y. Li, S. Zhang, L. Guo, J. Liu, Y. Wu, X. Wu. *NetNMSP: Nonoverlapping maximal sequential pattern mining*. Applied Intelligence, 52:9861-9884, 2022.

[24] Y. Wua, C. Zhu, Y. Li, L. Guo, X. Wue. *NetNCSP: Nonoverlapping closed sequential pattern mining*. Knowledge-Based Systems 196, 2020.

A Novel Big Data Intelligence Analytics Framework for 5G-Enabled IoT Healthcare

Yassine Sabri
Laboratory of Innovation in Management and Engineering for Enterprise (LIMIE), ISGA Rabat, Morocco

Ahmed Outzourhit
Laboratory of Innovation in Management and Engineering for Enterprise (LIMIE), ISGA Marrakech, Morocco

***Abstract*—Intelligent networking is a concept that enables 5G, the Internet of Things (IoT) and artificial intelligence (AI) to combine as a way to accelerate technological innovation and develop new revolutionary digital services. In the intelligent connectivity vision, the digital information gathered by the machines, devices and sensors which make up the IoT is analysed and contextualized. It is anticipated that the high availability of 5G and its inclusion of a large number of connections would help promote the production of wearable devices used to monitor the different biometric parameters of the wearer. Since these are AI-based health systems, the data obtained from these devices will be analysed in order to assess a patient's current health status. This paper presents a detailed design for the development of intelligent data analytics and mobile computer-assisted healthcare systems.The proposed advanced PoS consensus algorithm provides better performance than other existing algorithms.**

***Keywords*—Big data analytics; 5G-enabled; IoT healthcare; fog computing; confidentiality**

I. Introduction

According to the HIS market, by 2035, the Fifth Generation (5G) of wireless transmission technology will enable more than $1 trillion worth of products and services for the healthcare sector. The main features focused on in 5G technology, for example, significant increase in speed, coverage, enhanced capacity, network energy and power, and increased bandwidth, impact across many divisions in big data analytics. For people who have diabetes a comprehensive sensing analysis is available. Many mechanisms and personalized building models using 5G smart diabetes testing of smart clothing and smart monitoring using smartphones and using big data clouds are suggested for patients as part of their personalized solution for diabetes monitoring in healthcare. Here, an overall comprehensive process regarding blockchain-based 5G-enabled IoT and also various challenges and integration with blockchain industrial automation with the 5G-enabled IoT are presented. In addition, existing gaps in scalability, interoperability and other challenges in 5G blockchain are discussed. The deficiencies in 5G from all the communication devices and drones and particularly in the field of healthcare are identified and these problems will be overcome with the help of ultra-high reliability.

A. Motivation and Scope

The scope of 5G definition relates to potential uses and how those likely affect healthcare. The Internet of Medical Things (IoMT) focuses on the impact of 5G on providers, hospital systems, medical device companies, pharmacy companies and telehealth. Several key companies plan to launch 5G systems as well as 5G wireless networks in healthcare products. In the long term, in the healthcare sector, 5G will help to profoundly transform remote diagnostics and consultations. 5G enables the IoT to have a more powerful bandwidth combined with lower latency; 5G will be the main focus of technologies such as Augmented Reality and Virtual Reality in the healthcare sector. Furthermore, 5G will allow widespread deployment of Artificial Intelligence (AI) which will transform the healthcare sector from manual to smart automation.

Blockchain and 5G are the most hyped technologies emerging in the common marketplace. As mentioned, several features are available in blockchain with 5G, including decentralized approach, immutability, allows localized availability, cost efficiency and security. Also, challenges regarding blockchain with 5G integration focus mostly on scalability of blockchain which needs improvement to deal with the high number of devices and each device must have a unique address. Furthermore, after the 5G technology is deployed worldwide, it is expected that the technology will allow medical professionals to be able to exchange data with patients instantaneously from anywhere. It is an easy way for hospital-like monitoring in patient's homes similar to how intensive care units are monitored nowadays. Blockchain technology is perhaps the silver bullet needed for industry. The blockchain functions as a distributed transaction ledger for various IoT transactions. The blockchain platform supports and uses simple key management systems. The Ethereum platform is capable of managing a more fine-grained way used in many IoT devices with successful smart Turing complete code.

B. Research Contribution

In this paper, we proposed a 5G-enabled blockchain e-healthcare framework. The focus of this framework is a patient e-health management system. E-Health introduced the fog/edge used for easy access of medical data, as well as patient safety and privacy concern. Blockchain is deployed in the e-health system. This consist of three interfaces (1) Near Processing Layer, (2) Far Processing Layer and (3) Data Sensing Layer, and an agent Migration Handler (MH) used to monitor and transfer tasks which will help to locate the client. The current healthcare system is not patient friendly because patients must continually spend time monitoring their illness instead of resting, which is inconvenient for the patients. Wasted patient time has been reduced in our proposed system.

C. Organization

In Section 2 the literature survey and comparison of existing ideas with 5G technology in healthcare and blockchain is presented, in Section 3 our idea is proposed in a detailed manner, in Section 4 the performance analysis of the proposed model with graphs is discussed, and conclusions and future work in the healthcare sector are presented in Section 5.

II. Related Work

Some studies discussed by Hossain et al. proposed an emotion detection methodology in the healthcare system. They used IoT devices to capture emotion images and speech recognition processes separately, and calculated the value of the detected emotion and validated it [1]. Latif et al. discussed the 5G wireless technology along with emerging technologies that will transform the healthcare system, specifically 5G with cloud computing and 5G with artificial intelligence, and in terms of economist and high potential pitfalls in development of the 5G health revolution [2]. Similarly, Nasri et al. proposed a smart mobile IoT healthcare system using 5G and smart phones to monitor patient's health risk factors. WBSN data was used to monitor and track patient pulse, temperature and oxygen in blood as well as other vital parameters of the patients [3].

Ahad et al. discussed diabetes diagnosis with the solution of comprehensive sensing analysis. They suggested patients could have personalized solutions for diabetes monitoring in healthcare, including many mechanisms and personalized building models using 5G smart diabetes testing on smart clothing and smart monitoring using smartphones and big data clouds [4].

Further, Mistry et al. [5] presented a comprehensive review on blockchain-based 5G-enabled IoT and various challenges stemming from integration with blockchain industrial automation and the 5G-enabled IoT. A comparison of existing gaps between the scalability, interoperability and other challenges in 5G blockchain was also presented. Ullah et al. discussed the driving with wireless industry and developing the next generation of technology so that mobile technology generation has improved facilities to be efficient in wireless fields. Vehicle-to-everything (V2X) will impact in 5G with all the communication and drones and particularly in the field of healthcare, and they identified deficiencies and overcame those problem with the help of ultra-high reliability [6].

Furthermore, Li discussed how the next generation of wireless remote technology will be useful for healthcare in existing models with respect to the expenses of healthcare services and the imbalance of medical resources and inefficient healthcare system administration. To overcome this, the IoT, big data analysis,artificial intelligence technology and 5G wireless are used to improve patient quality of healthcare service, and the cost inferable method is focused on [7].

Sigwele et al. proposed an information and communication technology utilizing IoT to limit medical errors and cost of healthcare. They discussed the IEE5GG with smartphone gateway connection to save energy, which is executed with the help of MEC while considering QoS and battery level CPU load, and resulting with an energy efficient framework [8].

Chen et al. proposed a 5G-C-sys for healthcare that aims at ultra-low latency in cognitive application and high reliability. They also developed a prototype platform for 5G-C-sys incorporated with speech recognition and emotion detection for the effectiveness in healthcare-based 5G C-sys technology [9].

Similarly, Boban et al. analysed the requirement in 5G technology and identified the gaps with the existing technologies. They overcame the challenges with drone and communication technologies [10].

Latif et al. discussed how 5G technologies, AI, IoT and Big data will revolutionize healthcare, and they provided an overview of how machine learning algorithms are integrated and able to detect the anomalies in the healthcare system. The authors also investigated remote consultation in e-health [11]. Lakshmanan et al. proposed a hybrid approach in combining PSO and ACO & BCO on routing protocol and applying the K-Means algorithm for clustering the nodes [12].

Furthermore, Manoj et al. [13] discussed a congestion adaptive navigation for emergency situations. They also used sensors for locating using GPS and then server takes an action using PIR in emergency areas. Logeswari discusses the analysis of packets having the fuzzy logic based on the greedy routing protocol. Two characteristics input metrics and fuzzy decision making system in VANETs were used [14].

Gomathi et al. [15] proposed an energy efficient routing protocol using wireless sensors with dynamic clustering UWSN routing technique. This will be helpful for researchers due to reduced power consumption, response time, avoids overload and improves throughput of the network. Vignesh et al. discussed fewer deployments in the cloud storage with low cost replication, higher availability and better performance in geo-replicated systems by data centres with these benefits [16].

Ishwarya proposed a project to reduce congestion in traffic and calculated current traffic with normal. If anything unusual is detected, then emergency vehicles pass through the signals; thus, solving the traffic problems [17]. Sivasangari et al. [18] compared security and privacy using fog computing. They also used the fog computing principle to use a smart gateway for an improved big data health monitoring system. Suganthi et al. discussed security improvement for web based banking and the authentication using fingerprints to avoid hacking or for fault detection [19]. Deepa et al. proposed an idea of detecting road damage by image processing in smart phones and sending the co-ordinate point to the cloud and from the cloud a user can see the road where the damages are because it will show on a map. From this, they can avoid accidents and so on [20]. Keerthi et al. [21] used convolutional neural networks (CNN) to identify dangerous lung disease tumours. The CNN technique has many features and can provide standard representation pneumonic radiological complexity, fluctuation and classification of lung nodule. Sivasangari et al. [22] proposed major concerns about WBAN regarding the security and privacy of the healthcare sector. The patient health data should reach the physician at the right time. Security has the greatest impact on the lives of humans, and an effective model SEKBAN that ensures security data based on ECG signal was implemented. Indira et al. [23] implemented an efficient hybrid detection using a wireless sensor network. Wireless devices are spatially distributed over

sensors and physical changes. The device network includes multiple detection over sensors with lightweight transport.

Tao et al. [24] discussed V2G technology for enabling renewable energy sources providing power in a smart grid. They proposed a fog and cloud hybrid model. Vilalta et al. compares the existing approach with the new proposed technologies and distributed field. This paper discussed TelcoFog's benefits and dynamic deployment with low latency, and managing orchestration architecture for TelcoFog service infrastructure [25]. Furthermore, Chaudhary et al. focused on challenges for future demands in integrated fog computing and cloud computing in the 5G environment, in collaboration with SDN, NFV and NSC. They also performed data analytics with device mobility, as well as discussed challenges and potential attacks on the data shared in 5G [26].

Ku et al. [27] discussed advances in the fog radio access network and fog–cloud based in hybrid system issues. GPP is used for communication and computational processing, and it is also used as a simulator for experimental tests. Furthermore, Yang et al. [28] proposed an SDN-enabled approach for cloud–fog interoperation in 5G, and aimed at quality of optimized network usage. Crosby et al. [29] shared, in terms of blockchain technology, all criteria that satisfy specific application in both sectors regarding finance. There are many opportunities for revolution in disruptive technology. The digital currency Bitcoin is highly controversial, but blockchain has proved to be useful and has found many applications.

Risius et al. discussed a framework in blockchain which they divided into three group activities and four level of analysis. This paper addresses research predominantly focused on new technologies in blockchain [30].

Dinh et al. proposed a survey about untangling blockchain data processing with its challenges, and they analysed four dimensions in production as well as research systems—distributed ledger, smart contract, cryptography and consensus protocol. They also conducted comprehensive evaluation for major systems such as BlockBench, Parity and Ethereum, and found blockchain performance closer to the database [31].

Huh et al. [32-45] discussed how to manage IoT devices using blockchain to easily control and configure IoT devices. They also used the RSA algorithm which is capable of managing devices with secret keys. They used Ethereum blockchain for coding in an efficient way [33-37].

III. Proposed Work

Fog/Edge computing is tackled by eHealth systems for an easy way of accessing and processing medical data, and ensuring the privacy and safety of patients. All Fog and Cloud have their administers interested in handling medical data that violates the privacy of patients. Blockchain deployed over Fog and Cloud will allow patient data processing and storage. The eHealth architecture consists of three interface levels: sensing, near processing, and far processing. Multiple instances of a Patient Manager include 3-level structures. The Agent Migration Handler (MH) uses Profile Monitoring to transfer a task or execute the task internally, which collects profile information from remote agents.

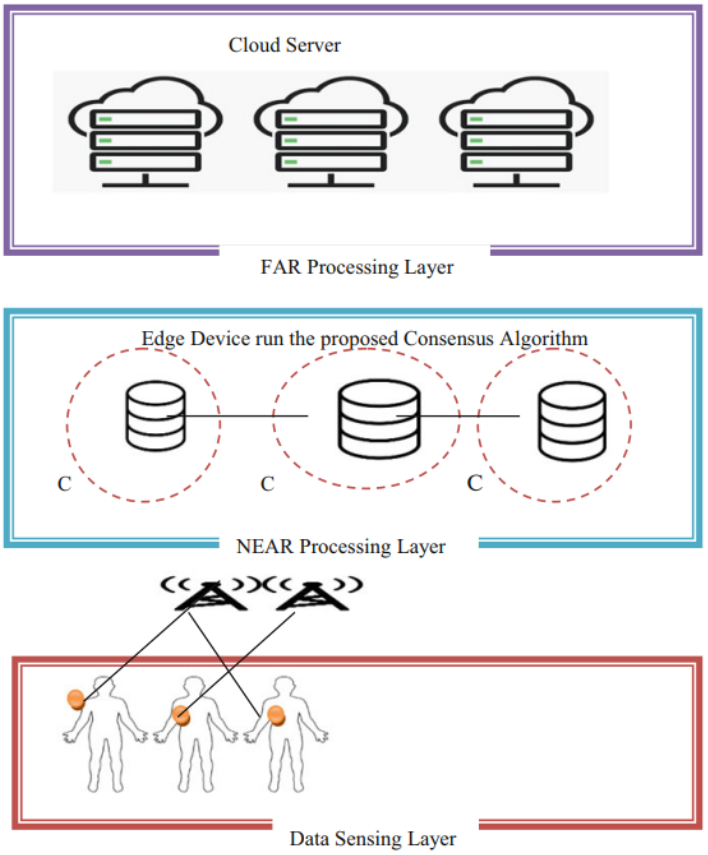


Fig. 1. Proposed Architecture.

Sensing layer: Smart devices, implantable sensors, smart watches, mobile devices and other devices monitor patients' body parameters. These applications use Bluetooth or ZigBee to convey to the mobile the physiological sign of a patient. Near processing layer: A hop away from the sensing devices of the data is where near processing level devices are generally located. Conventional switches, routers and low-profile devices are involved in the near networking layer. In a broad range of formats, healthcare facilities generate vast quantities of data, such as records, financial statements, laboratory findings, imaging tests, such as X-rays and CAD scans, and measurements of vital signs. Blockchain provides the ability to boost the data's authentication and integrity. This also helps to disseminate data inside the network or facilities. Such features have an effect on the cost, quality of data and reliability of providing health care across the system. Blockchain is an open, decentralized, intermediary-removing network. The blockchain healthcare solution does not require multiple authentication levels and provides everyone who is part of the blockchain architecture with access to the data. Data is made open and transparent for customers. Such apps will help to tackle the various issues facing the healthcare industry today. In the healthcare sector, blockchain's role is split into four stages. The proposed architecture is explained in Fig. 1.

The inspiration behind blockchain and 5G integration largely stems from the many benefits of blockchain in addressing security, protection, networking and service management issues in 5G networks. The proposed advanced Pos Consensus

algorithm is described below:

Algorithm 1 The Proposed Advanced Pos Consensus Algorithm

Input: Performance Transaction (PT), Reputation Transaction (RT), Stake Transaction, Agent Number (Ni) in a cluster

Output: Every fog agent generates $\mathrm{PTi}, \mathrm{STi}$ and produce RTi from the service provider

Form clusters with fog nodes within a threshold range (R)
for each cluster $k = 1 \in l$ **do**
  while Head election $=$ true **do**
    for member agent $i = 1 \in n_k$ of a cluster **do**
      Extract parameters from PT_i, RT_i to produce P_i, R_i and S_i
      $P_i = \frac{1}{1+e^{-\frac{\ddot{Y}}{r \times \alpha_i}}}$
      $R_i = \frac{1}{1+e^{-r}}$
      $S_i = \frac{1}{1+e^{-c}}$
      $f_i <-$ Decision Tree (P_i, S_i, R_i)
      $T_i = \Delta T \times \left(1 - \frac{f_i}{\sum_n^1 f_i}\right)$
      Every Member node in the cluster sets their timer (T_i)
    end for
    if (T_i) is expired **then**
      Then broadcast node id to the cluster for approval
    end if
    if approval count [node id] $>= 2/3 \times n_k$ **then**
      leader $_j <-$ nodeid
    end if
  end while
end for

A cluster within the Near Processing Layer of a certain geographic range (R). A cluster is formed by a fog/edge agent with a different patient member value, where a representative is selected to be the head of the cluster. The cluster head (also called the leader) is involved in running the blockchain consensus protocol by locking a certain amount of stake in the network. From each cluster, a cluster head (CH) is chosen, taking into account the member nodes' multi-criteria. The selection process includes the performance characteristics of a node, its reputation and the stake amount. Criteria are combined to measure a fitness value using a decision tree. The blockchain records the information of each node regarding the parameters listed, and can be retrieved from the blockchain. The performance parameters include device processing speed, storage capabilities, accessibility, variation distance coefficient and delay in transmission of an Agent. Here, MIPS processing capacities, memory space and availability are symbolized respectively as p1, p2 and p3.

T is the time interval for the selection of cluster head, and where T represents a limited random time period used to distinguish waiting time for the same fitness of the Agent. The Agent broadcasts its identifier across the cluster since its waiting time expires. The other cluster members verify the estimated fitness of the Agent and accept their approval for this Agent. In turn, every node in a cluster will participate in the PoS proposed. This consensus mechanism would be less vulnerable to an attack of 51% from each cluster than DPoS as a leader. The rich node, such as PoS, is less likely to become a cluster leader, as the cluster head is not only selected based on the locked coin.

Decision Tree is a supervised learning method which can be used for problems with classification and regression, but is preferred to solve problems with classification. It is a tree-structured classification where even the internal nodes represent the characteristics of a dataset, the branches represent the rules of decision and each leaf node represents the result.

IV. Performance Analysis

The assessment of the proposed programme with simulation settings and evaluation metrics is defined in this section. It also addresses the effects of various parameters such as energy usage and time for block generation. Various scenarios with different configurations are visualized through graph plot. For the following parameters, the performance of the updated mechanism and the existing mechanism will be investigated. Energy consumption: energy consumption refers to the energy needed for transmission, receipt of the transaction and simulated validation of the network of a number of blocks.

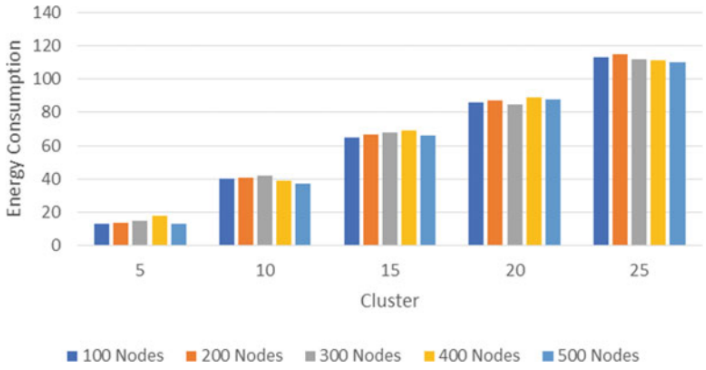


Fig. 2. Energy Consumption vs Cluster.

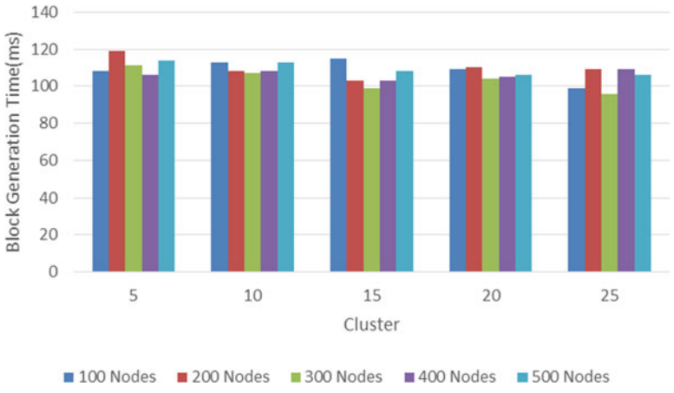


Fig. 3. Cluster vs Throughput.

Block generation time: This refers to the time required for a certain number of simulated network blocks to be uploaded, constructed and validated. In the simulated network, the updated process is executed five times and the output graphs are shown with average values generated from 10 execution runs. The regular one runs on a horizontal network and is supposed to function on a hierarchical network with the modified one. For both forms of consensus structures, nodes that lock digital coins into the network engage in mining. The energy consumption and execution time necessary for the development of 100 blocks are shown in Fig. 2, provided that the variable number of nodes and clusters is taken into account.

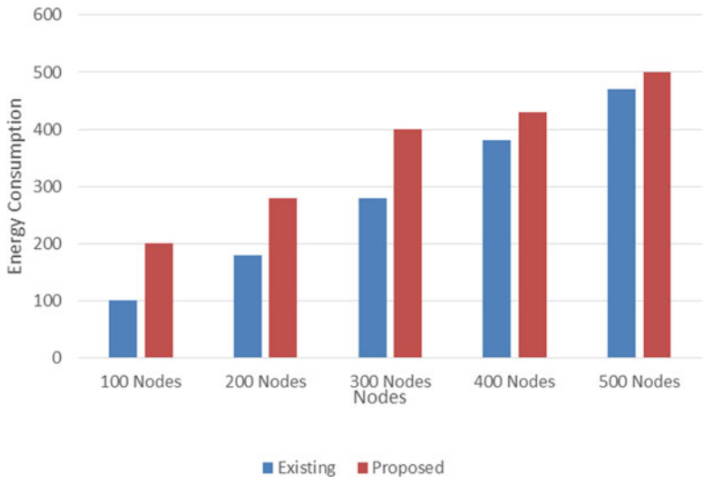


Fig. 4. Energy Consumption vs Nodes.

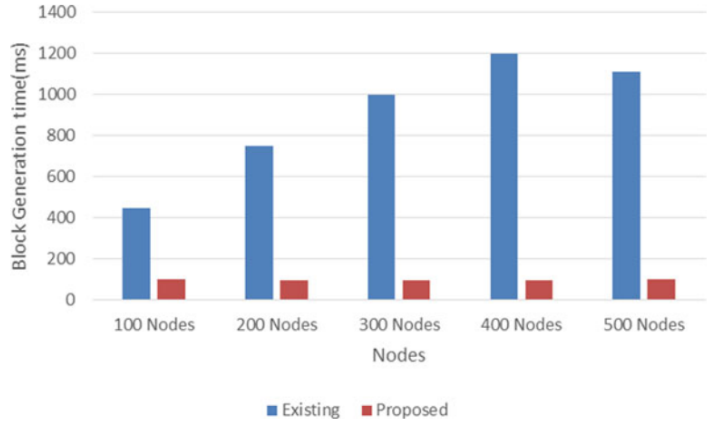


Fig. 5. Node vs Block Generation Time.

For unique clusters and nodes, the block generation time is shown in Fig. 3. The illustrated graph in Fig. 3 indicates that a pattern which is consistently lower or higher does not follow the period of block generation with a larger number of clusters. With a higher number of clusters, cluster heads collect transactions and construct blocks, with a higher number of blocks per second being generated. On the other hand, because of the delay in testing blocks, higher block generation time was also noticed for some higher cluster numbers.

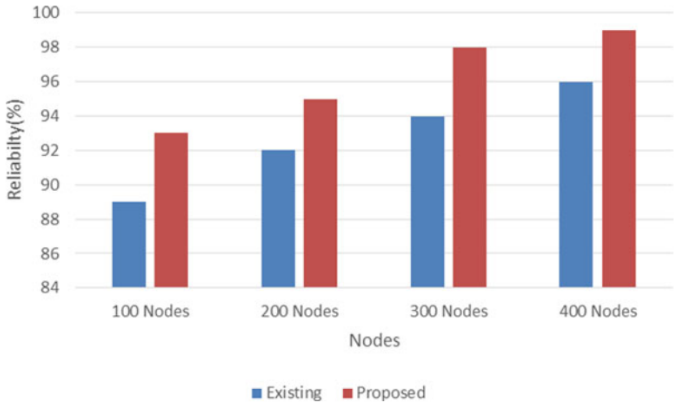


Fig. 6. Reliability vs Nodes.

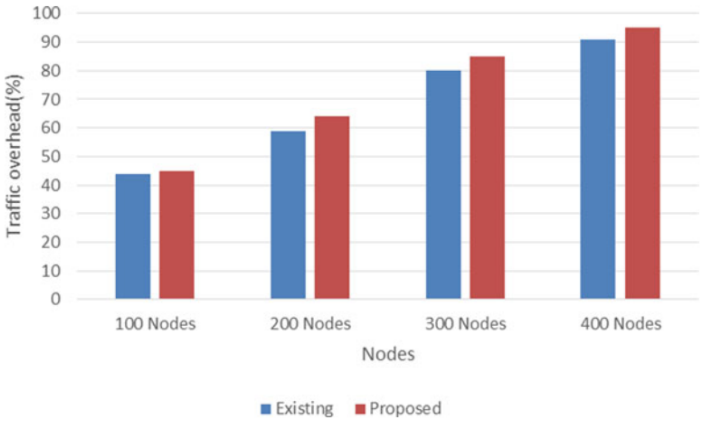


Fig. 7. Nodes vs Traffic Overhead.

In terms of power consumption and block generation time, the performance of the modified algorithm is compared to the standard one around. The revised algorithm shows a significant decrease in energy consumption compared to the regular one. A block is validated by a few selected safe miners in the updated one, but the standard one requires more than 50% node participation in the block validation process, resulting in higher energy consumption. Updated energy consumption remains almost constant for a comparable number of clusters with a greater number of nodes, while energy consumption tends to increase as the number of nodes within the network increases (Fig. 4).

Fig. 5 displays the updated and standard block generation period. The graph in Fig. 5 illustrates that the time for the standard generation of blocks is greater than for the updated one. As normal, different nodes send their transactions to one leader node for validation, and to broadcast across the network, a validated block is required. The approach thus consumes higher energy and makes it possible to take a longer time for the block's network-wide casting. In addition, some good miners are selected based on reputation, results and stake in the revised one, but a miner based on investment or stake alone is regularly nominated. We have painted

our architecture with an already proven architecture in terms of reliability and overhead touch. The protection protocol is correlated with these two performance metrics. The graph in Fig. 6 shows that our eHealth is more robust than the current system due to our decentralized Key Management and several three layer Patient Agent instances.

On the other hand, the diagram shown in Fig. 7 showed that our eHealth security mechanism provided greater overhead communication than the current one. An Agent needs a certain number of data encryption segments to be obtained from other neighbouring Agents to form the entire secret key. This technique activates overhead communication when exchanging hidden keys and authenticating. The relation between the different features and the current system [22] is shown in Table I. Similar to the cloud, with different protection strategies or without protection, different stakeholders deploy heterogeneous fog devices. Fog networks, through the identification and analysis of health information, are vulnerable to malicious attacks. In our architecture, the same patient agent replicated in the Mobile, Fog scheme and Cloud will protect wellbeing. To keep a Record of Malicious Attacks, sensitive medical information is analysed in the homogeneous replicated Patient Agent in order to protect the privacy or confidentiality of the patient.

A dropping assault occurs if a cluster head reduces the transactions. This is unlikely to happen because the cluster head will lose its reputation to share when it is detected as malicious. The consensus mechanism should select the malicious and cluster members who do not collect transactions for verification while the head of the cluster is down.

TABLE I. COMPARISON WITH EXISTING SYSTEM

Criteria	Proposed	Existing system
Confidentiality, availability and integrity	CAI high, PA homogeneous Responsive medical data task Edge nodes, using the ring signature, blockchain maintenance for metadata ensures several PAs Availability of service	Privacy is average, integrity is high, usability is high, Low because of centralized operations of blockchain Controller
Secure and energy efficient migration	High	Low
Communication overhead	High	Medium
Consensus mechanism	Light weight	Medium

V. CONCLUSION

In this paper, we built an eHealth program that deployed several instances of a three-layer Patient Agent software: sensing, near processing and far processing layer, which make the eHealth software more stable and fault-sensitive. We also defined how to implement the Patient Agent on a 5G unit. The dedicated Patient Agent application is able to handle the resources of 5G network slices. A performance analysis has shown that the emerging eHealth program will use blockchain technology to process health data in near-real time. The implementation of blockchain healthcare technology is difficult, with vast volumes of health data constantly being transmitted from wearable sensors.

REFERENCES

[1] M.S. Hossain, G. Muhammad, Emotion-aware connected healthcare big data towards 5G. IEEE Internet Things J. 5(4), 2399–2406 (2017)

[2] S. Latif, J. Qadir, S. Farooq, M.A. Imran, How 5g wireless (and concomitant technologies) will revolutionize healthcare? Future Internet 9(4), 93 (2017)

[3] F. Nasri, A. Mtibaa, Smart mobile healthcare system based on WBSN and 5G. Int. J. Adv. Comput. Sci. Appl. (IJACSA) 8(10), 147–156 (2017)

[4] A. Ahad, M. Tahir, K.-L.A. Yau, 5G-based smart healthcare network: Architecture, taxonomy, challenges and future research directions. IEEE Access 7, 100747–100762 (2019)

[5] I. Mistry, S. Tanwar, S. Tyagi, N. Kumar, Blockchain for 5G-enabled IoT for industrial automation: A systematic review, solutions, and challenges. Mech. Syst. Signal Process. 135, 106382 (2020)

[6] H. Ullah, N.G. Nair, A. Moore, C. Nugent, P. Muschamp, M. Cuevas, 5G communication: An overview of vehicle-to-everything, drones, and healthcare use-cases. IEEE Access 7, 37251–37268 (2019)

[7] D. Li, 5G and intelligence medicine—How the next generation of wireless technology will reconstruct healthcare? Precis. Clin. Med. 2(4), 205–208 (2019)

[8] T. Sigwele, H. Yim Fun, M. Ali, J. Hou, M. Susanto, H. Fitriawan, Intelligent and energy efficient mobile smartphone gateway for healthcare smart devices based on 5G, in 2018 IEEE Global Communications Conference (GLOBECOM), (IEEE, 2018), pp. 1–7

[9] M. Chen, J. Yang, Y. Hao, S. Mao, K. Hwang, A 5G cognitive system for healthcare. Big Data Cogn. Comput. 1(1), 2 (2017)

[10] M. Boban, A. Kousaridas, K. Manolakis, J. Eichinger, W. Xu, Connected roads of the future: Use cases, requirements, and design considerations for vehicle-to-everything communications. IEEE Veh. Technol. Mag. 13(3), 110–123 (2018)

[11] S. Latif, J. Qadir, S. Farooq, M.A. Imran, How 5G (and concomitant technologies) will revolutionize healthcare. Future Internet 9(4), 1–10 (2017)

[12] L. Lakshmanan, A. Jesudoss, V. Ulagamuthalvi, Cluster based routing scheme for heterogeneous nodes in WSN–A genetic approach, in International Conference on Intelligent Data Communication Technologies and Internet of Things, (Springer, Cham, 2018), pp. 1013–1022

[13] S.M. Kumar, L. Lakshmanan, A situation emergency building navigation disaster system using wireless sensor networks, in 2018 International Conference on Communication and Signal Processing (ICCSP), (IEEE, 2018), pp. 378–382

[14] K. Logeshwari, L. Lakshmanan, Authenticated anonymous secure on demand routing protocol in VANET (Vehicular adhoc network), in 2017 International Conference on Information Communication and Embedded Systems (ICICES), (IEEE, 2017), pp. 1–7

[15] R.M. Gomathi, J.M.L. Manickam, A. Sivasangari, P. Ajitha, Energy efficient dynamic clustering routing protocol in underwater wireless sensor networks. Int. J. Netw. Virtual Organ. 22(4), 415–432 (2020)

[16] R. Vignesh, D. Deepa, P. Anitha, S. Divya, S. Roobini, Dynamic enforcement of causal consistency for a geo-replicated cloud storage system. Int. J. Electr. Eng. Technol. 11(3), 181– 185 (2020)

[17] M.V. Ishwarya, D. Deepa, S. Hemalatha, A. Venkata Sai Nynesh, A. PrudhviTej, Gridlock surveillance and management system. J. Comput. Theor. Nanosci. 16(8), 3281–3284 (2019)

[18] A. Sivasangari, P. Ajitha, E. Brumancia, L. Sujihelen, G. Rajesh, Data security and privacy functions in fog computing for healthcare 4.0, in Fog Computing for Healthcare 4.0 Environments, ed. by S. Tanwar, (Springer, Cham, 2021), pp. 337–354

[19] D.S. Sharmila, L. Lakshmanan, Security improvement for web based banking authentication by utilizing fingerprint. Glob. J. Pure Appl. Math. 13(9), 4397–4404 (2017)

[20] D. Deepa, R. Vignesh, A. Sivasangari, S.C. Mana, B. Keerthi Samhitha, J. Jose, Visualizing road damage by monitoring system in cloud. Int. J. Electr. Eng. Technol. 11(4), 191–203 (2020)

[21] B. Keerthi Samhitha, S.C. Mana, J. Jose, R. Vignesh, D. Deepa, Prediction of lung cancer using convolutional neural network (CNN). Int. J. Adv. Trends Comput. Sci. Eng. 9(3), 3361–3365 (2020)

[22] A. Sivasangari, P. Ajitha, R.M. Gomathi, Light weight security scheme in wireless body area sensor network using logistic chaotic scheme. Int. J. Netw. Virtual Organ. 22(4), 433–444 (2020)

[23] K. Indira, D.U. Nandini, A. Sivasangari, An efficient hybrid intrusion detection system for wireless sensor networks. Int. J. Pure Appl. Math. 119(7), 539–556 (2018)

[24] M. Tao, K. Ota, M. Dong, Foud: Integrating fog and cloud for 5G-enabled V2G networks. IEEE Netw. 31(2), 8–13 (2017)

[25] R. Vilalta, V. López, A. Giorgetti, S. Peng, V. Orsini, L. Velasco, R. Serral-Gracia, et al., TelcoFog: A unified flexible fog and cloud computing architecture for 5G networks. IEEE Commun. Mag. 55(8), 36–43 (2017)

[26] R. Chaudhary, N. Kumar, S. Zeadally, Network service chaining in fog and cloud computing for the 5G environment: Data management and security challenges. IEEE Commun. Mag. 55(11), 114–122 (2017)

[27] Y.-J. Ku, D.-Y. Lin, C.-F. Lee, P.-J. Hsieh, H.-Y. Wei, C.-T. Chou, A.-C. Pang, 5G radio access network design with the fog paradigm: Confluence of communications and computing. IEEE Commun. Mag. 55(4), 46–52 (2017)

[28] P. Yang, N. Zhang, Y. Bi, L. Yu, X.S. Shen, Catalyzing cloud-fog interoperation in 5G wireless networks: An SDN approach. IEEE Netw. 31(5), 14–20 (2017)

[29] M. Crosby, P. Pattanayak, S. Verma, V. Kalyanaraman, Blockchain technology: Beyond bitcoin. Appl. Innov. 2(6–10), 71 (2016)

[30] M. Risius, K. Spohrer, A blockchain research framework. Bus. Inform. Syst. Eng. 59(6), 385– 409 (2017)

[31] T.T.A. Dinh, R. Liu, M. Zhang, G. Chen, B.C. Ooi, J. Wang, Untangling blockchain: A data processing view of blockchain systems. IEEE Trans. Knowl. Data Eng. 30(7), 1366–1385 (2018)

[32] S. Huh, S. Cho, S. Kim, Managing IoT devices using blockchain platform, in 2017 19th International Conference on Advanced Communication Technology (ICACT), (IEEE, 2017), pp. 464–467

[33] Al-Namari, Marwan A., Ali Mohammed Mansoor, and Mohd Yamani Idna Idris. "A brief survey on 5G wireless mobile network." International Journal of Advanced Computer Science and Applications 8.11 (2017).

[34] A Nasri, Farah, and Abdellatif Mtibaa. "Smart mobile healthcare system based on WBSN and 5G." International Journal of Advanced Computer Science and Applications 8.10 (2017).

[35] Alenazi, Bayana, and Hala Eldaw Idris. "Wireless Intrusion and Attack Detection for 5G Networks using Deep Learning Techniques." International Journal of Advanced Computer Science and Applications 12.7 (2021).

[36] Tahir, Sabeen. "A novel architecture for 5G ultra dense heterogeneous cellular network." International Journal of Advanced Computer Science and Applications 9.11 (2018).

[37] Sabri, Yassine, Aouad Siham, and Aberrahim Maizate. "Internet of things (iot) based smart vehicle security and safety system." International Journal of Advanced Computer Science and Applications 12.4 (2021).

[38] Yassine SABRI, Aouad Siham and Aberrahim Maizate, "Internet of Things (IoT) based Smart Vehicle Security and Safety System" International Journal of Advanced Computer Science and Applications(IJACSA), 12(4), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120487.

[39] Yassine Sabri, Najib El Kamoun, and Fatima Lakrami. 2019. Investigation of Energy Efficient Routing Protocols in Wireless Sensor Networks on Variant Energy Models. In Proceedings of the 4th International Conference on Big Data and Internet of Things (BDIoT'19). Association for Computing Machinery, New York, NY, USA, Article 51, 1–5. https://doi.org/10.1145/3372938.3372989.

[40] Yassine, S., & El Kamoun, N. (2017). Attacks and Secure Geographic Routing in Wireless Sensor Networks. Indonesian Journal of Electrical Engineering and Computer Science, 5(1), 147-158.DOI: http://doi.org/10.11591/ijeecs.v5.i1.pp147-158.

[41] Sabri, Y. (2016). GRPW-MuS: Geographic Routing to Multiple Sinks in connected wireless sensor networks. Indonesian Journal of Electrical Engineering and Computer Science, 4(3), 486-498.

[42] Sabri, Y., & El Kamoun, N. (2012, December). A prototype for wireless sensor networks to the detection of forest fires in large-scale. In 2012 Next Generation Networks and Services (NGNS) (pp. 116-122). IEEE.

[43] Yassine, S. A. B. R. I. (2022). A Routing Protocol for The Wireless Body Area Sensor Network (WBASN). IAENG International Journal of Computer Science, 49(2).

[44] Yassine, S., & Fatima, L. (2019, October). Dynamic cluster head selection method for wireless sensor network for agricultural application of internet of things based fuzzy c-means clustering algorithm. In 2019 7th Mediterranean Congress of Telecommunications (CMT) (pp. 1-9). IEEE.

[45] Y. Sabri and N. El Kamoun, "A prototype for wireless sensor networks to the detection of forest fires in large-scale," 2012 Next Generation Networks and Services (NGNS), 2012, pp. 116-122, doi: 10.1109/NGNS.2012.6656065.

A New Learning to Rank Approach for Software Defect Prediction

Sara Al-omari
Department of Computer Science
Applied Science Private University
Amman, Jordan

Yousef Elsheikh
Department of Computer Science
Applied Science Private University
Amman, Jordan

Mohammed Azzeh
Department of Data Science
Princess Sumaya University for Technology
Amman, Jordan

***Abstract*—Software defect prediction is one of the most active research fields in software development. The outcome of defect prediction models provides a list of the most likely defect-prone modules that need a huge effort from quality assurance teams. It can also help project managers to effectively allocate limited resources to validating software products and invest more effort in defect-prone modules. As the size of software projects grows, error prediction models can play an important role in assisting developers and shortening the time it takes to create more reliable software products by ranking software modules based on their defects. Therefore, there is need a learning-to-rank approach that can prioritize and rank defective modules to reduce testing effort, cost, and time. In this paper, a new learning to rank approach was developed to help the QA team rank the most defect-prone modules using different regression models. The proposed approach was evaluated on a set of standardized datasets using well-known evaluation measures such as Fault-Percentile Average (FPA), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Cumulative Lift Chart (CLC). Also, our proposed approach was compared with some other regression models that are used for software defect prediction, such as Random Forest (RF), Logistic Regression (LR), Support Vector Regression (SVR), Zero Inflated Regression (ZIR), Zero Inflated Poisson (ZIP), and Negative Polynomial Regression (NPR). Based on the results, the measurement criteria were different than each other as there was a gap in the accuracy obtained for defects prediction due to the nature of the random data, and thus was higher for RF and SVR, as well as FPA achieved better results than MAE and RMSE in this research paper.**

***Keywords*—Software engineering; software testing; software defect prediction; learning to rank approach**

I. Introduction

Predicting the exact and precise defect number is the best and most accurate option for software engineers, but because of the difficulty of achieving this task in real scenarios, it is not enough to rely on classifying modules into defects or not, so there is a need for another solution that can improve the defect prediction performance and increase the quality assurance of confidence in defect prediction [1]. This solution can be achieved by using a learning to rank approach that supports defect prediction models to rank and prioritize modules based on certain factors [1].

The importance of SDP models for predicting software defects has been discussed by using the LTR approach to rank a program according to the number of defects. The new model is supposed to improve the performance of the existing defect prediction models. Predicting the number of defects in software modules using machine learning regression models. However, this paper proposes a new learning to rank approach that supports defect prediction models for ranking and prioritization.

Most of the research in the past decade has focused on proposing new indicators for constructing predictive models [1]. The most studied indicators are the source code and process metrics [2]. Source code metrics measure the complexity of the source code [2]. Process metrics are derived from software documents, such as version control systems and issue trackers, which regulate the entire development process. Process metrics quantify many aspects of the software development process, such as source code changes, ownership of source code files, and developer interaction. The process metrics used to predict errors have been validated in many studies [2].

Defect prediction research is generally based on machine learning [3]. Predictive models built using machine learning approaches can predict the probability of errors in the source code (classification) or the number of errors in the source code (regression). Some studies have proposed the latest machine learning techniques, such as: improving active learning and prediction. The researchers also focused on determining the accuracy of predictions. Failure prediction models attempt to identify failures at the system, component, package, or file/class level. According to recent research, errors in modules or methods can also be identified and changed to different levels. Better accuracy can help developers by limiting the scope of source code reviews ensure quality. Suggesting a pre-processing method for predictive model is also an important research put forward in error prediction research. Before building the model, the following methods can be used for prediction: function selection, normalization, and noise protection [3]. Through the proposed pre-processing method, the predictive characteristics of actions in related research can be improved [3]. The researchers also proposed methods to predict defects in software projects [3]. The majority of the above representative studies were performed and verified within an internal prognostic framework, and the predictive model was developed and tested within the same project [4]. However, this is difficult for new projects that lack development history information. Create a predictive model. Typical methods for predicting crossover errors are metric compensation [4], nearest neighbor (NN) filters, naive transfer Bayes (TNB), and TCA+ (state-of-the-art transfer learning approach). Adjust the predictive model by selecting similar instances, transforming data values,

or developing new models [4].

Another important topic for defect prediction between items is to study the possibility of cross-prediction. Several studies have confirmed that cross-prediction is difficult to achieve; only a few cross-prediction combinations are effective [5]. Determining cross-prediction capabilities will play a major role in predicting errors between projects. There are many studies on the possibility of cross-prediction based on decision trees [5]. However, their decision tree has only been tested on certain software datasets and has not been studied.

The purpose of the SDP for the classification task is to predict which modules are likely to contain the most defects in order to allocate efforts to improve software quality, which means relative prediction and extraction of the exact number of defects, but this requires many conditions and accurate data to give the exact number of defects, which becomes difficult when the data is too large. However, the Learning to Rank (LTR) method provides a linear model by directly improving classification performance. It has been verified that it is useful to make forward adjustments to the classification performance metrics of the SDP model for constructing classification problems [6].

In this paper, a new learning to rank approach was developed to help the QA team with the ranking of the most defect-prone modules. The proposed approach was evaluated on a set of benchmark datasets using known measures such as fault percentile average (FPA), cumulative lift chart (CLC), mean absolute error (MAE), and root mean square error (RMSE). Our proposed approach will be compared with the current learning to rank approaches used in defect prediction.

The paper is organized as follows: Section II presents work related to software defect prediction as well as learning to ranking methods. Section III presents the proposed model including the data sets as well as the evaluation metrics used. Section IV presents the implementations made in the paper and finally Section V presents the findings and discussions about them before ending with the paper's conclusion.

II. Related Work

A. Software Defect Prediction

There are many studies that address the issue of predicting software defects. Among them, for example, X. Huo and M. Li, in [4] who proposed a new perspective for software defect prediction. This clearly articulates the "pair-wise" relationship between the bad module and the clean module to better prioritize the modules that are prone to failure, thus using benchmark dataset to ensure software reliability. X. Jing et al. in [8] attempt to systematically summarize all the typical work on predicting software failures in recent years. Based on the results of this work, this paper will help software researchers and professionals to better understand previous failure prediction studies based on datasets, software indicators, scores, and technical modeling perspectives in a simple and effective way. A. Okutan and O. Yıldız in [9] used Bayesian networks to study the relationship between software performance and error propensity. They used 9 records in the Promise data repository and showed that RFC, LOC, and LOCQ are the most error prone. On the other hand, the effect of NOC and DIT on defects is limited and unreliable. Y. Ma et al. [10] looked at a cross-company defect prediction scenario in which the source and target data came from different companies. They presented a novel technique called Transfer Naive Bayes (TNB), which uses the information of all the proper features in training data to select training data that is similar to the test data. J. Zheng in [11] studied three cost-sensitive impulse approaches for driving neural networks to predict software failures using four datasets related to a single action from the NASA project. Experimental results show that threshold shift is the best choice for cost-effective prediction of software failures using neural network models from the three approaches studied, especially for project datasets developed in object-oriented languages. X. Jing et al. in [12] used vocabulary learning methods to predict software errors. They used the characteristics of open-source software measurement to study various vocabularies (including error-free modules and damaged modules and sub-words of general vocabulary) and sparse representation coefficients. The dataset from the NASA project is used as a benchmark for evaluating the performance of all comparison methods. Experimental results show that CDDL is superior to several typical current error prediction methods. G. Czibula et al. in [13] proposed a classification model based on the mining of relational association rules. It is a discovery of relational association rules that can be used to predict whether a software module is flawed or not. On the open-source NASA datasets, an experimental evaluation of the proposed model. The results reveal that the classifier outperforms existing machine learning-based defect prediction approaches for the majority of the assessment measures studied. I. Laradji et al. in [14] introduced a two-variant (with and without feature selection) ensemble learning technique that is robust to both data imbalance and feature redundancy. Poor characteristics do not affect ensemble learners like random forests and the proposed technique, average probability ensemble (APE), as much as they do weighted support vector machines (W-SVMs). Furthermore, for the NASA datasets PC2, PC4, and MC1, the APE model paired with greedy forward selection (improved APE) attained AUC values of roughly 1.0. S. Liu et al. [15] employed the FECAR feature selection framework with Feature Clustering and Feature Ranking to forecast software defects. Using the FF-Correlation metric, this framework divides original features into k clusters. Then, using the FC-Relevance measure, it selects relevant features from each cluster. The data is based on real-world projects such as Eclipse and NASA. P. Krause and N. Fenton in [16] Focuses on a model developed for the Philips Software Center (PSC) using the expertise of the Philips Research Laboratory, which is specifically designed to predict the number of errors in various testing and operational phases. Seven of the 28 projects can obtain comprehensive data (completed questionnaires, more project data, and more error data). The study was not as successful as expected, and the authors confirmed that more investigations will be conducted in the future. The research is still in progress.

Li, M. Shepperd, and Y. Guo in [17] investigated the use and performance of unsupervised learning techniques in predicting software defect by conducting a systematic literature review that identified 49 studies with 2456 individual experimental results that met our inclusion criteria and were published between January 2000 and March 2018. Everything is in order. In this study, unsupervised classifiers did not appear

to perform worse than supervised classifiers.

T. M. Khoshgoftaar and colleagues in [18] proposed a methodology that incorporates a feature selection approach for picking relevant qualities and a data sample approach for resolving class imbalance. They used nine software measurement datasets from the PROMISE software project repository. Experimental results show that feature selection based on sample data performs significantly better than feature selection based on raw data, and the fault prediction model can achieve the same effect whether the training data is sample data or raw data.

L. Son et al. in [19] proposed a methodological mapping where they dealt with nine studies questions similar to distinctive stages of improvement of a DeP model. They explored every issue related to the method from collecting records; Preprocess records, strategies used to build a DeP fashions for the metrics used to evaluate the overall performance of the model and statistical evaluation plans used to mathematically validate the results of the DeP model. Out of the full 156 research, they decided on ninety-eight research for addressing 9 studies questions fashioned for this systematic mapping.

M. Sohan et al. in [20] used a lot of project data to prepare a balance and unbalanced dataset to build a prediction of software defects. Experimental results show that no significant changes are observed between balanced and unbalanced learning models. For a balanced learning model with an unbalanced test dataset, only the AUC value (area under the curve) increases exponentially. X. Cai et al. in [21] proposed a hybrid multi-purpose dynamic local search Cuckoo Search (HMOCS) to simultaneously identify health solutions. The problem of class mismatch in the dataset and the selection of SVM (support vector machine) parameters is critical to the prediction software defect. Eight datasets were selected from the Promise database to verify the proposed model for predicting software failures. Compared with the results of 8 prediction models, this method effectively solves the problem of predicting software failure. W. Li et al. in [22] proposed a two-step classification method and a two-step classification method based on three-way decision-making to predict cost-sensitive software failures by using NASA data. On the same direction Abu-Alhija et al. [23] studied the impact of kernels and SVM on the performance of defect predictions. they found that RBF is more Superior than other kernels.

B. Learning to Rank Approaches in Software Defect Prediction

X. Yang et al. in [1] used the LTR methodology for a wide range of real-world datasets and provided a full evaluation and comparison SDP for the ranking job, which included 10 construction approaches compared to other approaches on eleven real-world datasets. The relationship between CLC and FPA was also explored, as well as the need for metric selection over two sets of data for SDP for the ranking assignment. For the ranking job, the LTR technique to building SDP models yielded good accuracy and clarity of understanding. Also Xiaoxing Yang et al. in [2] used the learning-to-rank approaches to anticipate software defects. They presented the experimental results, which include a comparison of their approach to three other approaches from the literature, as well as five publicly available datasets. They employed the evolutionary optimization method to directly optimize the model performance measure, fault-percentile-average, which is not the same as the loss functions. For most datasets, the proposed learning-to-rank approach outperformed linear regression and logistic regression in terms of fault percentile-average models. Z. Cao et al. in [3] employed a learning to rank based approach to address the lack of legacy specifications that quantify the possibility of a candidate rule becoming a specification using 38 interesting measures. The benchmark dataset contains 28 classes from the Java 6 SDK that have been manually identified as having specification rules. These guidelines were derived from the completion of 14 projects. Experimental results using classes from the Java 6 SDK show that our learning to rank-based technique can enhance the best ranking performance using a single measure by up to 66 percent. X. Yu et al. [5] investigated the effect of 23 learning to rank approaches for EADP using 41 releases of 11 open source software projects taken from the PROMISE data repository to examine the impact of 23 learning to rank techniques for EADP. When the 23 approaches are trained on the original feature subset, the experimental findings demonstrate that BRR performs best in terms of FPA, while BRR and LTR perform best in terms of Norm (Popt) subset.

M. Buchari Yu et al. in [6] used two public benchmark datasets to create and assess the implementation of Chaotic Gaussian Particle Swarm Optimization on the Learning-to-Rank software defect prediction methodology for train model parameter. They conclude that using Chaotic Gaussian Particle Swarm Optimization in a Learning-to-Rank strategy can increase defect module ranking accuracy in datasets with high-dimensional characteristics. Y. Ma et al. in [7] used a top-k learning to rank (LTR) approach in the scenario of CPDP. The PROMISE dataset shows that SMOTE-PENN outperforms the other six competitive resampling approaches and Rank Net performs the best for the proposed.

III. The Proposed Model

A. Dataset

In this paper, a benchmark dataset was used from several types of versions, and the dataset was collected from the GitHub libraries and from previous research. The dataset applied to the developed regression models was 28 in total with different features, as shown in Table I below. The methodology on which the dataset was applied is to read the required data, then it was ensured that the data did not contain null values, and then the data was divided into x (features) and y (total defect). Finally, feature scaling technique was applied to make the output the same standard as it mentioned in the implementation section.

B. Developing Regression Model

In this paper, a model was proposed to predict the defects of the software modules and then rank the most defect-prone modules using six regression models such as (Random Forest, Logistic Regression, Support Vector Regression, Negative Binomial Regression, Zero Inflated Regression, and Zero Inflated Poisson). However, after we prepared the dataset, we applied the data to our modules. They are divided into two categories: variations of the Poisson regression model and regression trees:

TABLE I. BENCHMARK DATASETS

Datasets	Feature Number	Total Defect
ant-1.7	20	746
camel-1.0	20	339
camel-1.6	20	965
data_arc	20	225
data_ivy-2.0	20	352
data_prop-6	20	644
data_redaktor	20	175
JDT_R2_0	48	2397
JDT_R2_1	48	2743
JDT_R3_0	48	3420
JDT_R3_1	48	3883
JDT_R3_2	48	2234
jedit-3.2	20	272
jedit-4.2	20	367
log4j-1.1	20	109
lucene-2.0	20	195
PDE_R2_0	48	576
PDE_R2_1	48	761
PDE_R3_0	48	881
PDE_R3_1	48	1108
PDE_R3_2	48	1351
poi-2.0	20	314
synapse-1.0	20	157
synapse-1.2	20	256
velocity-1.6	20	229
xalan-2.4	20	723
xerces-1.2	20	440
xerces-1.3	20	454

1) Variations of Poisson Regression Model: NBR (Negative Binomial Regression) has been commonly used for SDP. ZIP, ZIR, and NPR are all variations of Poisson regression. When the response variable of the dataset contains a large number of zeros, the Poisson regression model will reduce the probability of zeros. Zero-inflated models can explicitly model the excessive occurrence of zero faults. Zero-inflated models assume that zero-defect modules come from two distinct sources.
2) Regression Trees: SVR, LR, and RF are different types of regression trees. RF is an ensemble classifier consisting of many trees, and outputs the average of individual trees .LR Random Forests is a set of decision trees that have been combined to form an ensemble. It is an approach for Supervised Learning. Several decision trees are used to process the input data. It is powered by constructing a variable number of decision trees at training time and displaying the class that is the mode of the classes or mean prediction (for regression) of the individual trees. SVR attempts to predict actual values. To separate the data, this technique employs hyperplanes. If this separation is not achievable, the kernel trick is used, in which the dimension is increased, and the data points become separable by a hyperplane. Logistic regression is a data analysis technique that is used to define and explain the connection between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

C. Evaluation Measures

In this paper, different measures were used to evaluate the accuracy of our modules, the goal of accuracy evaluation is to make it easier to determine which modules we evaluate is good. The evaluation method is to obtain the percentage. The percentage of defects in the preceding modules of the ranking is commonly applied. To evaluate SDP models for the ranking task. The following are the evaluation measures we used:

- Fault-Percentile-Average: FPA is one of the evaluation measures which could reflect the effectiveness of different prediction models across all cuts off values as shown in equation 1. FPA is the average of the proportions of actual defects in the top m (m=1,2,..,k) modules to the whole defects, which is a more comprehensive performance measure than the percentage of defects in the top 20% modules. A higher FPA means a better ranking, where the modules with most defects come first [1].

$$\frac{1}{k}\sum_{m=1}^{k}\frac{1}{n}\sum_{i=k-m+1}^{k}\frac{1}{n}n_i \tag{1}$$

  where:
  - k is the number of software modules.
  - n is the total number of defects in all modules.
  - m is the modules to the whole defects.

- Root Mean Square Error:
  RMSE stands for Root Mean Squared Error. The standard deviation of the errors that occur when making a prediction on a dataset is known as the RMSE. This is the same as MSE (Mean Squared Error), but the root of the number is considered when calculating the model's accuracy. The errors are squared before being averaged in RMSE as shown in equation 2. This means that RMSE gives larger mistakes a higher weight. This suggests that RMSE is far more beneficial when substantial errors exist and have a significant impact on the model's performance. This characteristic is important in many mathematical calculations since it avoids taking the absolute value of the error. In this metric as well, the lower the value, the better the model's performance.

$$RMSE = \sqrt{\left[\sum\left(P_i - O_i\right)/n\right]} \tag{2}$$

  where P_{i} is the predicted value for the i^{th} observation in the dataset. O_{i} is the observed value for the i^{th} observation in the dataset. n is the sample size.

- Mean Absolute Error:
  The Mean Absolute Error (MAE) is a statistic that assesses the average magnitude of errors in a set of predictions without taking their direction into account as shown in equation 3. The Mean Absolute Error is the average of the absolute differences between prediction and actual observation over the test sample, assuming that all individual deviations are equally weighted. It is less susceptible to outliers than MSE because it does not penalize large errors. When performance is measured using continuous variable data, it is commonly employed. It produces a linear value that equalizes the weighted individual disparities. The model's performance improves as the value decreases.

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} |o_i - P_i| \quad (3)$$

- Cumulative Lift Chart:
  A lift chart graphically Represents the improvement provided by the mining model to random estimation and measures the change in the form of elevation estimation as shown in equation 4. Through comparing the elevation estimates of different models, you can determine which model is better.

$$CLC = FBA - \left(\frac{1}{2k}\right) \quad (4)$$

  where k is the number of software modules.

D. Research Methodology

In this paper, a new learning to rank (LTR) approach was developed to help the QA team rank the most defect-prone modules in the software and thus reduce testing efforts using various regression models. The datasets used were taken from the standard dataset, and the datasets are divided into training and test data. In the Software Defect Prediction Program (SDP), training data and test data were selected in two separate ways. First, in the same dataset, the training and test data were randomly selected (or may be sequential). In the second stage, the training will be taken from the dataset as the previous version, and the test data from another dataset will be taken as the next version. The first approach was adopted and used. We then evaluated the data using known evaluation measures such as Fault-Percentile-Average (FPA), Mean Absolute Error (MAE), Root-Mean-Square Error (RMSE) and Cumulative Lift Chart (CLC). Our proposed LTR approach was compared with the current LTR approaches used in software defects prediction. Fig. 1 illustrates the research methodology used in this paper.

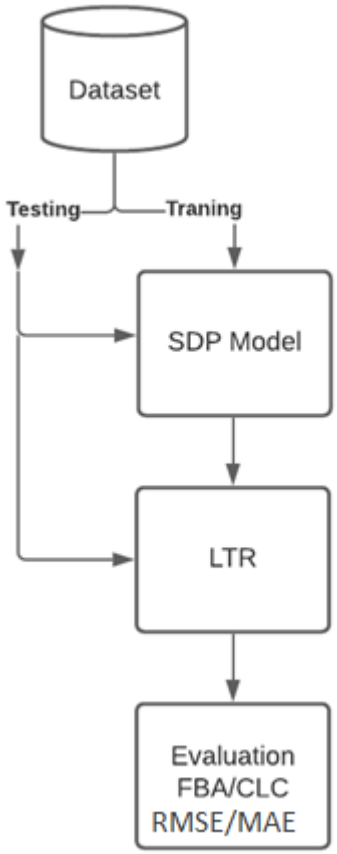


Fig. 1. Research Methodology.

E. Implementation

To prove the success of our proposed LTR approach, it is necessary to apply our work and show and compare the results. Various regression models were used in a separate way from previous studies, by applying the LTR approaches and the programming language that we will discuss. Python 3.6 and Spyder 3.2.6 were used to evaluate the accuracy of the ML regression models (SVM, RF, LR, ZIP, ZIR and NPR). In addition to the usage of Google Colab to run the existing LTR approaches to do comparison with the proposed LTR approach in this paper. Each model was developed separately from the others, but in this section, we collected the models to present the methodology in an obvious way. Each model was used from twenty-eight datasets. The databases were configured prior to use so that they were all applied in a uniform manner. We will go through the methodology in a clear manner by explaining the steps of the code.

IV. Result

In this paper, four evaluation measures were used to calculate the accuracy of our regression models, and we will present the findings in tables depending on the evaluation measures. Twenty-eight datasets were used over six regression models in this paper. The goal of this paper is to present the accuracy of our software modules by applying 5-fold cross-validation technique. The reason to use 5-fold cross-validation is to get the best result based on 28 datasets with different features.

A. Fault Percentile Average

By calculating the average accuracy of Fault-Percentile-Average using 5-fold cross validation, we show the values in Table II.

The goal of building these regression models was achieved by finding the model that contains the largest number of errors and comparing the machine-trained model with the outputs in the data. The accuracy was measured to discover the model that predicts the number of errors, and the accuracy expresses the result of the prediction of the machine in how close it is to the original result. Here, the closer the result is to zero, the better the result. From our experience, it is difficult to determine which model is better because of the amount of disparate data, but we can determine the best model by comparing the models on one dataset only.

Fault-Percentile-Average (FPA) evaluation measure was used based on previous studies, which were studied based on classification models, thus showed satisfactory results, but in

TABLE II. COMPARISON OF THE LTR APPROACH WITH SIX EXISTING REGRESSION MODELS OVER 28 DATASETS WITH ALL METRICS USING FPA MEASURE

Datasets	LR	RF	SVR	ZIP	ZFR	NBR
ant-1.7	0.72154	0.77648	0.77122	0.76151	0.76784	0.81054
camel-1.0	0.60842	0.640408	0.66202	0.61714	0.57756	0.45306
camel-1.6	0.57958	0.71205	0.69675	0.71383	0.63459	0.72812
data_arc	0.53375	0.65012	0.62578	0.54251	0.5341	0.50761
data_ivy-2.0	0.52781	0.67103	0.65993	0.61565	0.54292	0.70769
data_prop-6	0.56486	0.64195	0.66541	0.66563	0.55886	0.71364
data_redaktor	0.647809	0.75695	0.74704	0.68828	0.64923	0.76476
JDT_R2_0	0.65196	0.72185	0.70785	0.69471	0.71048	0.67293
JDT_R2_1	0.63748	0.77034	0.74373	0.76109	0.67657	0.75939
JDT_R3_0	0.64215	0.784	0.77129	0.76294	0.73085	0.76901
JDT_R3_1	0.62876	0.77124	0.77164	0.74924	0.70828	0.75553
JDT_R3_2	0.70679	0.6923	0.768605	0.72612	0.76279	0.76752
jedit-3.2	0.7021	0.82434	0.79438	0.80392	0.79859	0.79549
jedit-4.2	0.68124	0.8172	0.78171	0.73489	0.71828	0.84774
log4j-1.1	0.77168	0.7857	0.77079	0.72377	0.76497	0.78366
lucene-2.0	0.73887	0.7384	0.75203	0.74142	0.74499	0.75741
PDE_R2_0	0.64553	0.79867	0.80697	0.66434	0.7267	0.783902
PDE_R2_1	0.7051	0.7989	0.78326	0.64177	0.69192	0.75939
PDE_R3_0	0.70525	0.7603	0.73343	0.72478	0.74422	0.72364
PDE_R3_1	0.72154	0.7458	0.755806	0.72328	0.73999	0.694006
PDE_R3_2	0.63911	0.6912	0.67414	0.618407	0.65225	0.60929
poi-2.0	0.53461	0.6831	0.66588	0.64228	0.5689	0.56535
synapse-1.0	0.72382	0.6905	0.54816	0.51767	0.64681	0.61529
synapse-1.2	0.68296	0.7114	0.70046	0.67401	0.68626	0.69261
velocity-1.6	0.63761	0.7448	0.73101	0.70212	0.614424	0.70044
xalan-2.4	0.53136	0.7831	0.73722	0.66721	0.56662	0.75826
xerces-1.2	0.5083	0.7058	0.67377	0.64692	0.54765	0.57162
xerces-1.3	0.65305	0.7982	0.78974	0.79185	0.63956	0.824109

this paper, we applied it to six regression models on a larger scale, so that we used all available databases in the field of rank learning, and we have obtained satisfactory results. In this paper, we demonstrated the success of the error-percentage-mean scale. All results were not shown over-fitting on the result.

As seen in Table II, the columns represent all the datasets we used and the rows represent the regression models that we created, for example row ant-1.7 represents the first dataset to which Fault-Percentile-Average has been applied to show the accuracy results for the regression model that was built To determine the model that contains the largest number of program errors, and this accuracy represents the proximity of the learned data to the test data and here we find that the best reading for it is 0.81054, which represents the negative binomial regression model, and this result does not mean that it is the best model because it may depend on the nature of the data and the evaluation measure.

B. Mean Absolute Error

By calculating the average accuracy of Mean-Absolute-Error using 5-fold cross validation, the values shown in Table III.

Because we are using regression models in this paper, it is necessary to mention the measurement criteria for the regression, such as Mean Absolute Error. We have used the same methodology in building defect models that contain the largest number of errors and measuring the average accuracy of the models by using 5-fold cross-valuation on a twenty-eight dataset with all features. As shown in Table III, the accuracy results from using the evaluation of the Mean Absolute Error of the regression. It is clear that some results have exceeded the relevance because most of the datasets are intended for classification. However, satisfactory results were shown in some datasets. This does not mean that other models failed to show

TABLE III. COMPARISON OF THE LTR APPROACH WITH SIX EXISTING REGRESSION MODELS OVER 28 DATASETS WITH ALL METRICS USING MAE MEASURE

Datasets	LR	RF	SVR	ZIP	ZFR	NBR
ant-1.7	0.41879	0.6446	0.49934	1823.4	0.47655	1.17607
camel-1.0	0.05307	0.0962	0.11	96198	0.044205	1.0006
camel-1.6	0.53989	0.821	0.5827	0.9038	0.569209	1.8976
data_arc	0.14222	0.2289	0.2044	0.281	0.15111	8.3864
data_ivy-2.0	0.13062	0.2011	0.1915	0.2532	0.11931	0.93563
data_prop-6	0.10558	0.1989	0.1755	0.2164	0.103101	0.92528
data_redaktor	0.12	0.7564	0.1836	0.3213	0.12	0.94013
JDT_R2_0	1.60409	1.85684	1.5485	11.67	1.9147	12.9109
JDT_R2_1	0.8217	0.96404	0.8325	1.006	0.9085	4.59836
JDT_R3_0	1.3888	1.643	1.268	1.42	1.632	9.2229
JDT_R3_1	1.16981	1.65063	1.141	1.568	1.463	28.6809
JDT_R3_2	1.0069	1.037	1.003	1.194	1.111	5.76459
jedit-3.2	1.26168	1.5983	1.309	21.56	1.374	13.4099
jedit-4.2	0.30729	0.44233	0.357	66.56	0.3646	1.05466
log4j-1.1	0.67878	0.8246	0.7729	97.44	0.6958	1.2819
lucene-2.0	1.1846	1.4018	1.228	1.5778	1.344	1.7224
PDE_R2_0	0.41136	0.5771	0.4617	0.704	0.455	1.2279
PDE_R2_1	0.32592	0.45809	0.378	1.4671	0.3647	1.1646
PDE_R3_0	0.68745	0.9269	0.6769	1.103	0.7752	12.8876
PDE_R3_1	0.73639	1.025	0.7483	0.9479	0.9085	8.2126
PDE_R3_2	0.8815	1.016	0.8026	2.779	1.155	10.796
poi-2.0	0.15596	0.2296	0.2037	2.028	0.1356	1.412
synapse-1.0	0.15342	0.2271	0.2283	8.472	0.1489	1.117
synapse-1.2	0.49894	0.6402	0.536	0.7406	0.5739	1.01748
velocity-1.6	0.82647	1.468	0.8462	34.75	1.071	13.4941
xalan-2.4	0.23104	0.3194	0.2719	0.3883	0.2319	0.94382
xerces-1.2	0.28181	0.4473	0.3314	5.049	0.3331	1.0469
xerces-1.3	0.37284	0.5871	0.4693	0.6761	0.333	1.23983

accuracy in every way, but rather they showed satisfactory results according to the nature of the data.

C. Root Mean Square Error

Table IV shows the RMSE results based on different numbers of matrices. We applied 10 times 5-fold cross-validation over 28 datasets with all metrics. By calculating the average accuracy of Root-Mean-Square-Error using 5-fold cross validation, we show the values in Table IV.

This is another way to calculate mean precision with 5-fold validation using RMSE. As shown in Table IV, more than appropriate occurred in some of the data, and this is because the nature of the data is for classification and not for regression. However, we have achieved satisfactory results, and these results were mostly concentrated on two models (Linear Regression and support vector regression).

D. Cumulative Lift Chart

This is a way to evaluate the measure of our modules to show the relationship between two evaluation measures (FPA, MAE) in easy and effortless way by presenting the chart of all six modules we have used before as shown in Fig. 2. The charts show the performance of our regression models against other well known LTR methods algorithms [1-2] [3-10].

V. CONCLUSION

SDP models for ranking task manage testing resources more effectively by predicting which modules are likely to have more errors in the software program. SDP data is gathered by a variety of IT organizations and individuals, and it is noisy. As a result, estimating the number of errors per software

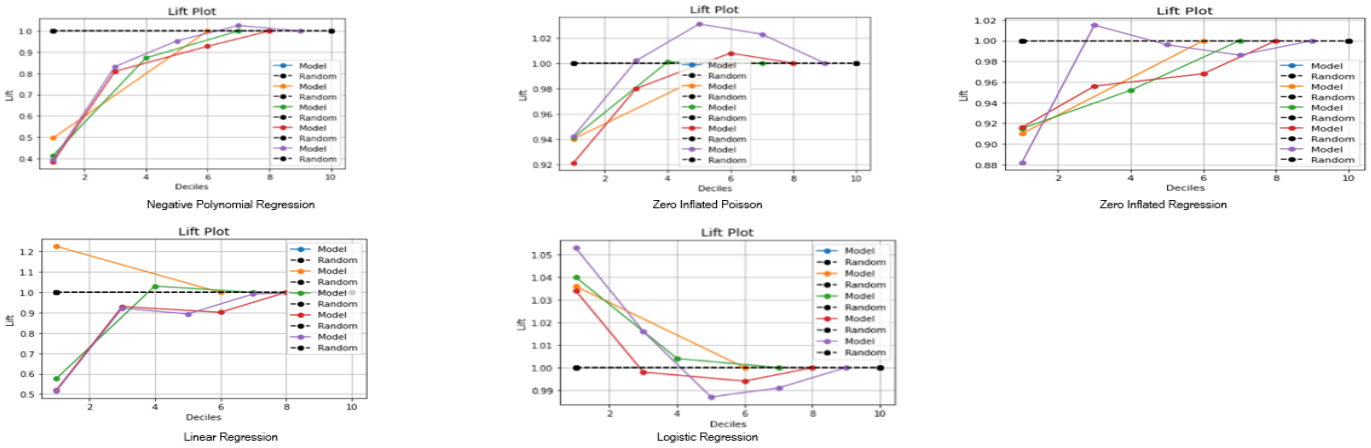


Fig. 2. Cumulative Lift Chart for Various Models.

TABLE IV. COMPARISON OF THE LTR APPROACH WITH SIX EXISTING REGRESSION MODELS OVER 28 DATASETS WITH ALL METRICS USING RMSE MEASURE

datasets	LR	RF	SVR	ZIP	ZFR	NBR
ant-1.7	0.6435	0.8016	0.70547	27.81	0.6898	1.0757
camel-1.0	0.2265	0.3097	0.3313	14872	0.2022	0
camel-1.6	0.728	0.9049	0.7573	0.936	0.7498	1.3337
data_arc	0.3754	0.478	0.4517	0.5301	0.3831	2.4535
data_ivy-2.0	0.3604	0.448	0.4373	0.5031	0.3447	0.9671
data_prop-6	0.3233	0.4459	0.4183	0.465	0.3201	0.9618
data_redaktor	0.339	0.1953	0.4281	0.5668	0.3431	0.9695
JDT_R2_0	1.2608	1.3495	1.243	2.465	1.38	3.244
JDT_R2_1	0.894	0.9771	0.905	0.9989	0.9497	1.7868
JDT_R3_0	1.151	1.2797	0.905	1.183	1.272	2.7136
JDT_R3_1	1.06	1.273	1.044	1.23	1.205	3.5289
JDT_R3_2	0.9956	1.09	0.9929	1.088	1.053	2.024
jedit-3.2	1.101	1.258	1.126	3.988	1.165	2.98
jedit-4.2	0.5475	0.66471	0.591	5.421	0.6002	1.026
log4j-1.1	0.8222	0.9069	0.8762	5.2293	0.8325	1.1205
lucene-2.0	1.085	1.1798	1.104	1.2416	1.1497	1.2979
PDE_R2_0	0.636	0.7588	0.675	0.8356	0.6715	1.1061
PDE_R2_1	0.5614	0.6731	0.6063	1.7129	0.5924	1.0776
PDE_R3_0	0.8213	0.9577	0.8175	1.035	0.876	2.4717
PDE_R3_1	0.8567	1.0085	0.8625	0.9721	0.9505	2.5657
PDE_R3_2	0.9381	1.004	0.8932	1.454	1.068	2.6754
poi-2.0	0.394	0.4767	0.4502	1.242	0.3654	1.1445
synapse-1.0	0.3737	0.4758	0.4744	2.232	0.3666	1.0521
synapse-1.2	0.6936	0.7973	0.7256	0.8583	0.7557	1.0084
velocity-1.6	0.9029	1.2028	0.9148	4.828	1.032	3.3796
xalan-2.4	0.4798	0.565	0.521	0.6228	0.48	0.9714
xerces-1.2	0.53	0.6662	0.5751	1.917	0.5754	1.022
xerces-1.3	0.6088	0.565	0.679	0.8178	0	1.1126

module is difficult, if not impossible, due to a lack of precise historical data. Some academics propose utilizing a ranking-based performance metric to assess SDP models such as CLC and FPA. However, contemporary SDP models have been enhanced to properly predict a specific number of errors. However, a decent model based on individual loss functions may fail to provide a satisfactory ranking. As a result, in this paper, we proposed a unique approach, distinct from earlier studies, for developing models by direct improvement of the ranking performance measurement. We applied the LTR approach to a wide range of real-world datasets in this paper and present a complete assessment and comparison of RF, SVR and LR with other approaches. We also estimated the error using FPA and MAE and then used CLC to explain the disparity between its results.

The following are the key findings from our research paper:

1) Employing the regression approach rather than the classification approach, as opposed to prior studies in the literature where the classification technique is employed. This is to highlight the contrast between the classification and regression models. Whether or not this model has mistaken, the data is divided into 0 and 1, with 0 containing no errors and 1 containing errors. This is known as classification. However, the regression models that we are working on estimate the number of mistakes in each website, which means that the first website based on the characteristics (x values) has a number of errors, and so the regression models train the model when the data enters it. The characteristics will predict mistakes that are either equal to or near to the amount of genuine errors. This is the point of using regression models.
2) Proposing a new LTR approach with scipy.stats and apply it to multiple models to compare and calculate accuracy. We discovered that the model produced using regression models accomplished what was expected of it in terms of identifying models with the highest number of errors, and the percentage of accuracy varied according to the type of data. And according to the comparison with the standard measures, we found that the model RF and SVR is better.

Based on the results and their comparison, we found that the measurement criteria differ from each other, so that we found a gap in calculating the accuracy in some measurements due to the nature of the random data, and FPA achieved better results than MAE, RMSE in this research paper.

ACKNOWLEDGMENT

Yousef Elsheikh and Sara Al-omari are grateful to the Applied Science Private University in Amman, Jordan, for the financial support granted to cover the publication fee of this research article.

Mohammad Azzeh thanks the Princess Sumaya University for Technology for supporting this research.

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] X. Yang, K. Tang and X. Yao, "A Learning-to-Rank Approach to Software Defect Prediction", IEEE Transactions on Reliability, vol. 64, no. 1, pp. 234-246, 2015. Available: https://www.ieee.org/publications/rights/index.html.

[3] Xiaoxing Yang, Ke Tang, and Xin Yao ,"A Learning-to-Rank Approach for Constructing Defect Prediction Models", IEEE, vol. 1, no. 64, p. 9, 2021. Available: http://file:///C:/Users/pc/Desktop/Learning

[4] Z. Cao, Y. Tian, T. Le and D. Lo, "Rule-based specification mining leveraging learning to rank", Automated Software Engineering, vol. 25, no. 3, pp. 501-530, 2018. Available: https://ink.library.smu.edu.sg/sis_research/3988/.

[5] X. Huo and M. Li, "On cost-effective software defect prediction: Classification or ranking?", Neurocomputing, vol. 363, pp. 339-350, 2019. Available: https://www.journals.elsevier.com/neurocomputin.

[6] X. Yu, K. Ebo Bennin, J. Liu, J. Wai Keung, X. Yin and Z. Xu, "An Empirical Study of Learning to Rank Techniques for Effort-Aware Defect Prediction", IEEE, p. 12, 2021. Available: 2019.

[7] M. Buchari, S. Mardiyanto and B. Hendradjaya, "Implementation of Chaotic Gaussian Particle Swarm Optimization for Optimize Learning-to-Rank Software Defect Prediction Model Construction", Journal of Physics: Conference Series, vol. 978, p. 012079, 2018. Available: 10.1088/1742-6596/978/1/012079.

[8] Y. Ma, "A Top-k Learning to Rank Approach to Cross-Project Software Defect Prediction", IEEE, p. 11, 2021. Available: 2018.

[9] Z. Li, X. Jing and X. Zhu, "Progress on approaches to software defect prediction", IET Software, vol. 12, no. 3, pp. 161-175, 2018. Available: 10.1049/iet-sen.2017.0148.

[10] A. Okutan and O. Yıldız, "Software defect prediction using Bayesian networks", Empirical Software Engineering, vol. 19, no. 1, pp. 154-181, 2012. Available: 10.1007/s10664-012-9218-8.

[11] Y. Ma, G. Luo, X. Zeng and A. Chen, "Transfer learning for cross-company software defect prediction", Information and Software Technology, vol. 54, no. 3, pp. 248-256, 2012. Available: 10.1016/j.infsof.2011.09.007.

[12] J. Zheng, "Cost-sensitive boosting neural networks for software defect prediction", Expert Systems with Applications, vol. 37, no. 6, pp. 4537-4543, 2010. Available: 10.1016/j.eswa.2009.12.056.

[13] X. Jing, S. Ying, Z. Zhang, S. Wu and J. Liu, "Dictionary Learning Based Software Defect Prediction", p. 10, 2014.

[14] G. Czibula, Z. Marian and I. Czibula, "Software defect prediction using relational association rule mining", Information Sciences, vol. 264, pp. 260-278, 2014. Available: 10.1016/j.ins.2013.12.031.

[15] I. Laradji, M. Alshayeb and L. Ghouti, "Software defect prediction using ensemble learning on selected features", Information and Software Technology, vol. 58, pp. 388-402, 2015. Available: 10.1016/j.infsof.2014.07.005.

[16] S. Liu, X. Chen, W. Liu, J. Chen, Q. Gu and D. Chen, "FECAR: A Feature Selection Framework for Software Defect Prediction", IEEE, p. 10, 2014.

[17] P. Krause and N. Fenton, "A probabilistic model for software Defect Prediction", IEEE, p. 36, 2001.

[18] N. Li, M. Shepperd and Y. Guo, "A systematic review of unsupervised learning techniques for software defect prediction", Information and Software Technology, vol. 122, p. 106287, 2020. Available: 10.1016/j.infsof.2020.106287.

[19] T. M. Khoshgoftaar, K. Gao† and N. Seliya, "Attribute Selection and Imbalanced Data: Problems in Software Defect Prediction", IEEE, vol. 1, p. 8, 2010.

[20] L. Son, N. Pritam, M. Khari, R. Kumar, P. Phuong and P. Thong, "Empirical Study of Software Defect Prediction: A Systematic Mapping", Symmetry, vol. 11, no. 2, p. 212, 2019. Available: 10.3390/sym11020212.

[21] M. Sohan, M. Jabiullah, S. Motiur Rahman and S. Mahmud, "Assessing the Effect of Imbalanced Learning on Cross-project Software Defect Prediction", IEEE, 2019.

[22] X. Cai et al., "An under-sampled software defect prediction method based on hybrid multi-objective cuckoo search", Concurrency and Computation: Practice and Experience, vol. 32, no. 5, 2019. Available: 10.1002/cpe.5478.

[23] Al-Haija, Haneen Abu, Mohammad Azzeh, and Fadi Almasalha. "Software Defect Prediction Using Support Vector Machine." International Journal of Systematic Innovation 7, no. 2 (2022): 37-47.

Utilizing Artificial Intelligence Techniques for Assisting Visually Impaired People: A Personal AI-based Assistive Application

Samah Alhazmi, Mohammed Kutbi, Soha Alhelaly, Ulfat Dawood, Reem Felemban, and Entisar Alaslani
College of Computing and Informatics
Saudi Electronic University
Riyadh, Saudi Arabia 11673

***Abstract*—Nowadays, the Artificial Intelligence (AI) field has made a significant change in the real life. Numerous applications use the AI techniques for the purpose of assisting people in different life aspects. Furthermore, with the increased number of people with visual difficulties around the world, there is a need for such AI assistive applications which provide them an independent life. Limited affordable and appropriate solutions developed so far. In this paper, we present a personal AI-based assistive application called (Vivid) that supports visually impaired people being more independent. Vivid has many features such as identifying objects, objects' colors, recognizing text, and faces detection. It relies on using the mobile camera to sense the environment, and the machine learning techniques to understand the environment. By translating a meaningful information in audible sound for those users, Vivid does not require to have any visual ability. Moreover, the whole interaction with the user is only based on voice commands. The input from the user is captured as finger gestures on tablet or cell phone touch screen. In addition to Vivid, we also shade the lights on a supplementary application that notify/alarm visually impaired people of any nearby objects using sensors. These personal assistive applications were developed then tested on the real world and showed promising results.**

Keywords*—Artificial intelligence; machine learning; assistive technology; visually impaired

I. Introduction

Globally, there are around 285 million people who are considered visually impaired. Thirty-nine million of them are blind and the rest have considerably low visions abilities [1]. In the US, there were four million visual impairments cases in 2010, and are projected to be seven million cases and thirteen million cases in 2030 and 2050, respectively (NIH-NEI). Those people can benefit greatly and improve their life-independently by using AI-based assistive solutions. Even though many smartphone applications developed, limited applications focused on visually impaired people. It is very difficult for a blind or visually impaired user to use a smartphone effortlessly. However, there are several features that can allow those special users to utilize such technology seamlessly as regular users.

In this paper, we propose an AI-based assistive application (Vivid) for those targeted users which is affordable, accessible, and easy to use. Many of the required assistive features were combined such as: (1) colors identifier, (2) objects labels, (3) text reader, (4) facial expression, and (5) distance notifier. The first four features were implemented together in single camera-based application which allow the user to use (Vivid) without any assistance. The use of this application is based merely on finger gestures as an input, and voice feedback as an output. The user interface was made incredibly simple to provide seamless user experience for targeted users. Whereas the last feature "distance-notifier" was developed as a supplementary assistive application with Vivid. The reasons of separating these features in two different applications are: (1) reducing complexity; (2) the first application "Vivid" can be used totally by the blind person without the need for an assistant help, while the second application "distance-notifier", the user might need assistance from someone else to avoid wireless connection errors; (3) the user might only need to use the features provided by "Vivid" which are camera-based, thus, they only download "Vivid", which will provide more flexibility. Table I shows a brief description of the five features.

II. Related Work

The development of modern technologies helped to make all these technologies accessible by all categories of people. Modern technologies are not limited to be used by normal visually people only, however, they can be used by blind people as well. A few years ago, smartphones have been widely used by community – by normal people – and became most popular which touch our daily lives. Whereas, for visually impaired and blind people, technologies are still limited for them; however, the new technologies and smart solutions provided by smartphones encourage blind people to be more independent and self-reliance completely. Authors in [3], proposed a system that is based on Morse-Code, which is a code in which letters were represented by combinations of long and short light or sound signals.

Researchers in [4] raised a question in their study, how do blind users use smartphones? Usually, it depends on a screen reader that exists in its operating system, such as, Google operating system which known as Android, or Apple's system which known as IOS. But more than that, it depends on the presence of some other services, such as screen magnification and the development of night mode feature to suit some other visual disabilities. Moreover, there are other settings for disabled people within "General" menu under the name of "Accessibility". This feature allows disable people to choose what is suitable to them. In addition, smartphones have a feature which called "screen reader" to help in reading what is shown in the screen.

TABLE I. A BRIEF DESCRIPTION THE FIVE FEATURES OF THE VIVID APPLICATION

Feature	Description
Colors Identifier	To identify colors of any object by capturing an image of that object. Basically, the implemented algorithm detects the most dominant RGB colors, and identifies colors values; then, converts these values to colors names that are known by human. At the end, voice feedback is generated saying the identified color's name. This feature is not only assisted people with blindness, but also helps people who suffer from color-blindness.
Objects Labels	A machine learning tool, which is 'MLT kit' API [2] was used. The ML Kit mobile SDK brings all the Google's expertise on machine learning techniques to mobile developers in an easy-to-use package, our application can allow the user to identify any object name/label by taking an image of that object. For example, there is a "ball" object which the blind person cannot identify, by taking an image of that object, the application can identify it and tell the user it is a ball.
Text Reader	Again, by using the 'ML kit' API [2],this feature allows visually impaired people to be able to detect any written text and hear it; not only the text that is in braille. The user can take an image of the text by bringing the camera closer to where the text is written and taking an image of that text. The application then generates a voice version of that text.
Facial Expression	The 'ML kit' API [1] is used, this feature allows visually impaired people to be able to detect if there is a person in front of them with the use of face detector and then facial expression detector. User can take an image of the person in front of him/her by bringing the camera closer to where of his/her moving direction and take an image. The application then generates a voice version of that text.
Distance Notifier	To use this feature, the user will need a hardware sensor. The sensor is attached to the user as a belt and will notify/alarm him/her of any object that is getting closer or might hit him/her. The sensor used was Ultrasonic Sensor HC-S04, which is a hardware for identifying distances. The way how this hardware works, was by sending sound waves from the transmitter, which then, bounce off an object, and return to the receiver; the user can determine how far away something is, by the time it takes for the sound waves to get back to the sensor. To connect this sensor to the smartphone, we developed a separate application called "Distance-Notifier" that can handle wireless Bluetooth connection to this sensor and send alarms/notifications to the user.

Another study [5] discussed how smartphones can be used by blind people. In IOS, for example, it provides feature of screen reader known as "Voice Over", and it supports most of the languages in the world. When this option is activated, the use of this device is entirely different; the device turns into a speaker device. Any touch on the screen, will tell the user what is that touched point. For example, if the user touched an application icon i.e. (Facebook), the screen reader will say "Facebook twice to open" and will say this sentence in other languages as well. Next, the application is opened only after pressing it twice – this is similar to double-click the mouse when using a computer to move from one item to another. The screen is swiped from the left to right to go to the next item or application, or also to go to the previous item or application. With each swipe, a screen reader will utter the name of the item or application and the mechanism of its activation. The device is easy to use and depends completely upon gestures with fingers on the screen; it seems at the beginning a difficult job for blind people to use it, but it is actual requires some training to use it [5]. While IOS devices provide the service of Voice Over, Android devices which are widely used by people around the world provide the service of "Talkback" which is an accessibility service. It allows visually impaired people to interact with smartphones and use them regularly as everyone else does. It is based on spoken words, and other audible feedback that will give the user a full experience of what they are doing, and what the output produced by the device [6].

Authors in [7] said that we still convinced that modern technologies allow disable people to live in all life aspects with a realist and an effective manner. This enhanced the continuous development of modern technologies for those people. As discussed in [8], there were many text reader applications, such as, the scanner through the smartphone camera, and the conversion of written texts to audio. Moreover, authors in [9] implemented an Optical Character Recognition (OCR) program, which provides the opportunity to scan books and letters. The program works once any text has scanned, then, reads the text loudly. The OCR consists of a camera that captures the text, which then converted to speech through the program. The drawback of this program is that, this technique required a hardware in order to work sufficiently; which made it difficult to be available to everyone. However, by involving this technique into smartphones, it becomes more useful and usable [9].

Furthermore, an application developed by researchers in [10] which was based on two hardware devices. One for text input and the other for speech output. Basically, it works using a sensor component - as an eye - that captures any printed text. Then, it extracts the recognized text area, and produced a speech output - by the audio output device - for blind users. The main drawback of this application was the cost. Another study [11] showed the automation of text-to-audio. Basically, a pen-like device is used to convert any non-Braille text to audio. Any piece of text that a person would like to read, is converted to an audio signal; after that, by using the Bluetooth technology, these audio signals are transmitted to Bluetooth earphones. The authors in [11] believed that pen technology is lighter in the sense of it can be easily portable. This can change blind people life by allowing them to read whatever they desire.

In the International Conference on Computer and Information Technology [12], an Android system assistant for visually impaired people was developed which called Eye Mate which can help users to know where the obstacle is through vibrations. This android application provides navigation to a blind person and track his/her movement as well. This was based on a voice command; the application will generate a voice command according to the obstacle object position. The movement of a person are measured by using GPS which tracks the user position latitude and longitude. Furthermore, in [13] a sensor-based assistive device for visually impaired people was proposed. This device has a sensor to identify the distance between the person and a harmful object. Other studies [14][15] showed some devices that have been used by visually impaired people, however, these devices are big and not comfortable to use.

In [16], a proposed product called "Self-Energized Smart Vision Stick" developed for blind people, as shown in Fig. 1; the stick uses Arduino Ultrasonic Sensor. It is basically a sticklike a tool that provides safety and privacy to those people. This stick is attached to distance sensors to identify the distance between the stick and any nearby objects, then,

notify the user accordingly.

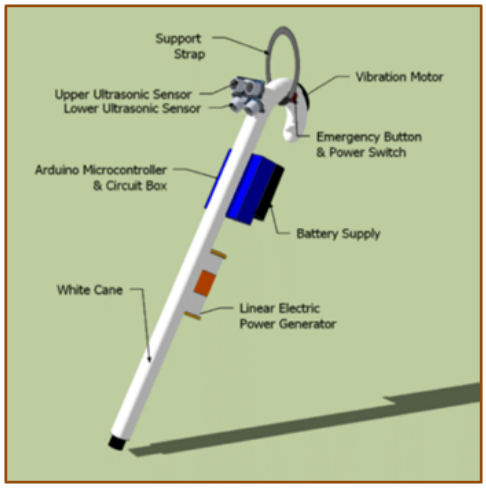


Fig. 1. General 3D Model of the Self-Energized Smart Vision Stick [16].

Authors in [17] declared that colors are a concern to blind people because it is a very important to them to know what the colors of their clothes are. There are many color-identifier applications proposed, such as "Color Reader" sensor. An American foundation for blind people invented a "color teller" an easy-to-use device, which helps visually impaired people to identify the color of any object. Moreover, the researchers in [18] highlighted the importance of colors, and implemented a device called "Coloresia". This device could convert any color into music or words. In addition, the American council for blind people [19] made a comparison between the color detectors applications in the market, and then, provided better solutions for these applications if needed. A list of other applications discussed that help visually impaired people in several ways, such as, a reader to identify colors and objects which work with QR technique [20].

Google Firebase is the newest technology of image labelling process. Image labelling provides information of what the images contain. With the use of ML kit API, it will enable the application to recognize objects within an image. Here, are some examples of objects that can be detected: people, activities, places, things, and so on. When it recognizes an object, it indicates a score of confidence level to show how confidently the machine learning could detect the object.

In [21], a video-based application was developed, where number of frames can be generated from the video, then, converting the images from RGB to Gray scale. This was done by applying ML algorithm, which set the key points of an object, then, match the object with the database object. If (object=database), then, convert the text to speech. Table II shows the advantages of "Vivid" as opposed to existing applications.

III. Materials and Methods

As stated in the previous section, our goal is to develop a comprehensive solution to help people with visual difficulties with low cost. Thus, we proposed a mobile application to assist visually impaired and/or color-blind people to be more independent by providing five features: Colors Identifier, Objects Labels, Text Reader, Facial expression, and Distance Notifier. We proposed two mobile applications which were linked together. The main application is "Vivid" and the second application is a "Distance-Notifier". The first application "Vivid" can be used totally by the blind person without the need for an assistant help, while the second application "distance-notifier", the user might need assistance from someone else to avoid wireless connection errors. The "Distance-Notifier" is connected to a hardware sensor that alarms/notifies the users of any nearby objects. We separated these features into two independent applications to reduce the complexity.

A. Vivid Application

Vivid is a camera-based application that captures objects/things and translates them into audible sound. Thus, considering our targeted users who will not be able to see the screen or to use the regular interface that consists of buttons and other controllers, Vivid was built with a simple interface. The features of colors identifier, objects labels, and anything reader are combined in single camera-based application that is called "Vivid". Thus, a blind/ visually impaired user can use "Vivid" application totally by himself/herself without any assistance because the use of this application is based merely on finger gestures for an input, and voice feedback for an output. The user interface was made incredibly simple to provide seamless user experience for blind/ visually impaired people.

The color identifier works over two steps: (1) find the dominant color D; (2), find what color is that. It estimate the dominant color by averaging the color of all pixels per channel $C_1, 2, 3(1 = red, 2 = green, 3 = blue)$:

$$C_i = \frac{\sum_{k=0}^{n} p_{\text{ik}}}{n} \tag{1}$$

where i is the channel number, k is the pixel index in the image and n is the total number of pixel in the image. The average RGB color is then refer to as the dominant color

$$D = (C_1,\ C_2, C_3) \tag{2}$$

After deciding the dominant color of the picture, RGB color is converted to HSV format. First, we get the maximum

$$C_{\max} = \max(C_1, C_2, C_3) \tag{3}$$

and minimum

$$C_{\min} = \min(C_1, C_2, C_3) \tag{4}$$

of the values from the three channels. Then,

$$a = C_{\max} - \ C_{\min} \tag{5}$$

Second, we calculate the value of Hue (H) and Saturation(S). H is calculated as follow:

$$H = \begin{cases} 0, & a = 0 \\ 60\ x\ (\ \frac{C_2 - C_3}{a} mod\ 6, & C_{\max} = C_1 \\ 60\ x\ (\ \frac{C_3 - C_1}{a} + 2, & C_{\max} = C_2 \\ 60\ x\ (\ \frac{C_1 - C_2}{a} + 4, & C_{\max} = C_3 \end{cases} \tag{6}$$

TABLE II. ADVANTAGES OF "VIVID" AS OPPOSED TO EXISTING APPLICATIONS

Existing Applications	Vivid
An Optical Character Recognition (OCR) program, which provides the opportunity to scan books and letters [9].	The mentioned technique requires a hardware in order to work sufficiently to convert text to audible sound, which made it difficult to be available to everyone. Vivid provides a solution to this drawback by involving this technique into smartphones. Since smartphones are accessible to almost everyone, getting access to this technology is only a few clicks away.
Portable camera based assistive text reading from handheld objects for blind person [10].	Vivid takes the advantage of already existing audio input/out means that are built in the smartphone. No additional hardware required. All the three tasks (scanning text, processing text, producing audio) are done instantaneously in one single device.
Automated electronic pen aiding visually impaired in reading, visualizing, and understanding textual contents [11].	The mentioned pen technology is light and portable. However, Vivid provides better advantages in a sense that despite of being available on a light, portable mobile phone, it is also accessible easily within a few clicks.
Android assistant EyeMate for blind and blind tracker [12].	The proposed technology uses GPS to track coordinates and notify of any obstacles. Vivid enhances this technology by using an Arduino sensor. The sensor measures the distance instead of using GPS system to provide better accuracy and more precise distance measurements.
Android assistant EyeMate [12].	In this application, the impaired person needs to deal with two different applications, (1) to capture contextual (distance of an obstacle, position of the sensors, environment around the user), and then, (2) to communicate with the other application to deliver this information to the user. Our Vivid's algorithm solved this issue by eliminating the need to communicate to any outsider. Thus, the algorithm itself handle the captured information and translate it to audible sound.
Optical devices for distance [14].	The technologies used by developing a device which made it too large and uncomfortable to use as compared to Vidid.
Self-energized smart vision stick for visually impaired people [15].	Using a sick that sense the nearby environment and obstacle is a smart idea. However, the stick is heavy, and the user needs to carry it around. Vivid provides a smart belt instead of a stick so that the user does not need to carry, they only need to wear it and the application will take care of notifying the user of any nearby obstacle.
Assistive technology products by the American Foundation for the blinds [16].	The foundation invented a "color teller" which is an easy-to-use device. It helps visually impaired people to identify the color of any object in front of them. However, Vivid is allowing the users to use the color detection services free of charge while the proposed device costs around 205 dollars.
Bilingual wearable assistive technology for visually impaired people [18].	The wearable hardware requires additional costs compared to Vivid which is merely a downloadable mobile application.
Color to sound converter for blind people [22]	The study suggests using a sensor called "Color Reader". Vivid on the other hand, does not require any additional sensors other than the camera of the mobile. The "color detection" algorithm integrated within Vivid uses the information captured by the camera and acts as a sensor that can identify colors.

Lastly, S is calculated using the following formula:

$$S = \begin{cases} 0\ , & C_{\max} \neq 0 \\ \frac{a}{C_{\max}}, & C_{\max} \neq 0 \end{cases} \tag{7}$$

Then, identifier takes the dominant color and classifies it to one of 12 predefined colors. Each of those colors has a range of numbers if the dominant color fill in the range, then it is classified with it. Table III shows the ranges of HSV values for color identification.Then,find what color is that by:

$$Color = f(H, S) \tag{8}$$

To make sure we are using state-of-the-art techniques, we used Android ML Kit called Firebase [2]. For object recognizer, it offers objects classification of 10000+ classes and 400 classes when working with the online version. The model's performance are measured in term of accuracy is the number of times the model correctly classifies an object. The accuracy for the model is 60% as reported by authors of ml kit library [2].

The model's architecture is MobileNet [23] to use minimal resources on the phone or tablet. For text reader, it recognizes the test in an image and split it into blocks which later read word-by-word out loud to the user. Lastly, facial expression recognizer, it has two features; first, it detects faces in the image then, second, it classifies the expression into smiling or not. The system output "There are no person" if there were no people detected in the captured image. If there is a person and its face is detected, it will classify the expression of the person. Then, it output the expression either smiling person or not smiling person.

Vivid application has a simple workflow which is described in Fig. 2. After launching the application, user manual will appear on the screen only if it is the first time the app has been launched. Otherwise, camera feed will be presented on the screen. Then, user press on the screen to capture specific picture of interest. Then, users have five core actions: swiping up to get object label recognition, swiping down for re-capturing picture of interest, swiping left for color detection, swiping right for text recognition and, finally, long-press for face expression.

Vivid application consists of two activities: the main activity and the image processing activity. The main activity as shown in Fig. 3(a) is the first interface the user will expect, which is responsible to capture the image. As shown in the figure below, the whole screen is merely the camera display, it would not be reasonable to display any sort of text output, graphic or buttons for user interaction with the app because the target users are visually impaired ones. Thus, user will interact by figure gestures. On this activity user can capture the desired object by bringing it close to the camera lens and then tabbing anywhere on the screen. Tabbing gesture will tell the camera to capture the scene at that moment and then send that captured content to the Image Processing activity for processing. However, the image processing activity as shown in Fig. 3(b) is responsible for processing the content captured in the previous activity. The interface again doesn't have any form of typical user interaction controls. Interaction is based on audible outputs and finger gestures. Once user hears the word "swipe" that indicate that the image is processed, and the application is ready to produce the output. User then can get the output using figure gestures of swiping in different directions to get different outputs as defined in the flowchart figure above.

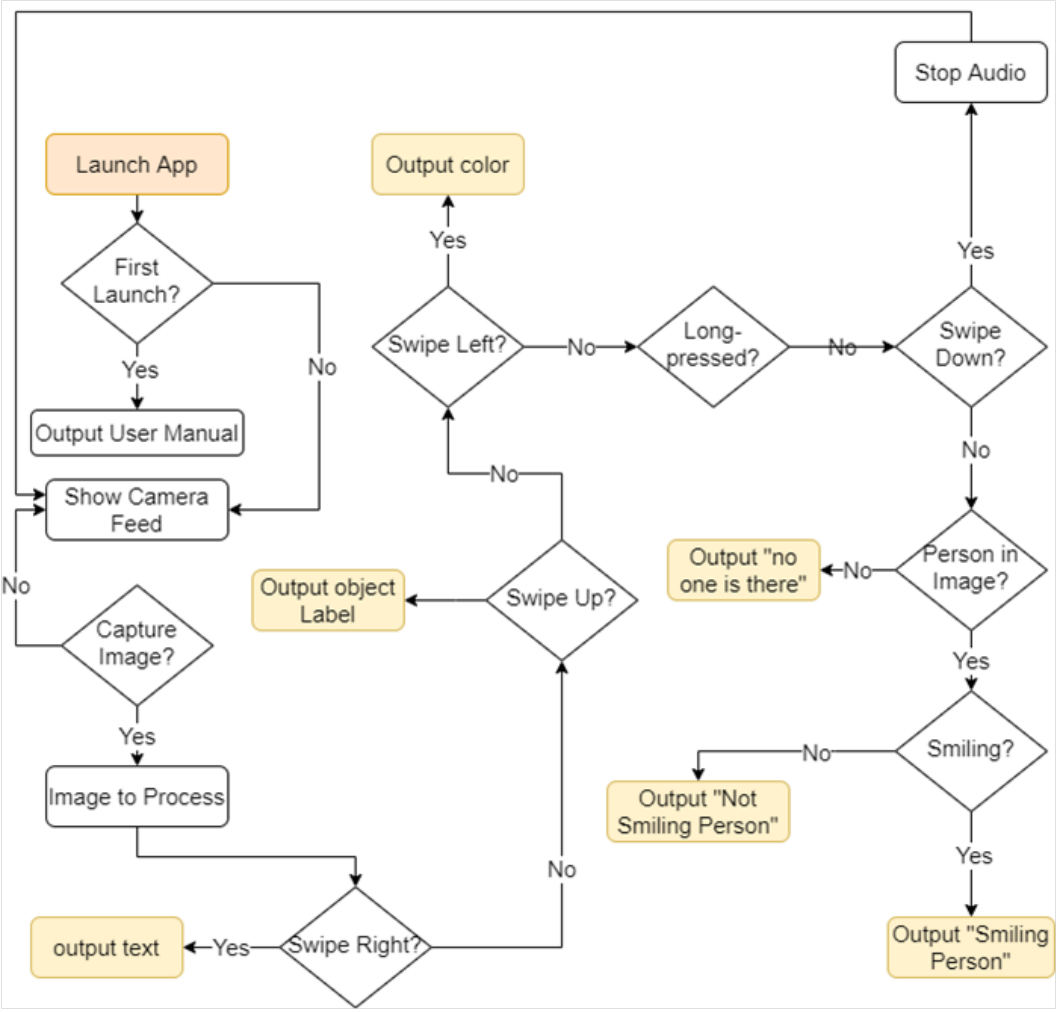


Fig. 2. Flowchart of Vivid Application. All Features and their Workflow of the Vivid Application is Explained in the Figure.

Fig. 3. Overview of the Interfaces of Vivid Application. (a) Main Activity. Camera Feed Showing before Any Interaction is Done by the User. (b) Image Processing Activity. After Taking a Picture.

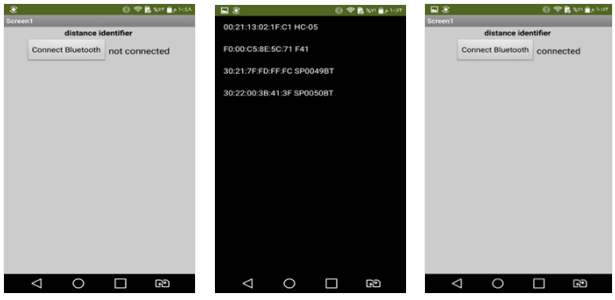
B. Distance-Notifier Application

Distance Notifier application uses a hardware sensor. The sensor is attached to the user as a belt and will notify/alarm him/her of any object that is getting closer or might hit him/her. The used sensor is Arduino Ultrasonic Sensor HC-S04, which is a hardware for identifying distances. The sensors would send sound waves from the transmitter, which then, bounce off of an object, and return to the receiver; you can determine how far away something is, by the time it takes for the sound waves to get back to the sensor. Therefore, to connect this sensor to the smart phone, we developed a separate application called “Distance-Notifier” that can handle wireless Bluetooth connection to this sensor and send alarms/notifications to the user. It is preferred that, to use an assistant person to setup and connect this application to the hardware, to avoid any connection errors. Afterwards, the blind user can receive notifications and alerts from this application.

Distance-Notifier application is a standalone which requires additional hardware. The additional hardware is Ultrasonic Sensor HC-SR04 and an Arduino [24]. The phone is connected to the Arduino using Bluetooth. The app receives the signals from the Arduino and output voice notification for the user. The voice notifications alert user about obstacles on her way. Fig. 4(a, b and c) shows in the interface of the distance notifier application. For the first interface we had a button that allows the user to select Bluetooth device, then, it will show to the user the list of Bluetooth devices, after the list, the user can select the desirable device. Finally, after selecting the device, the used connected and it will be ready to be used.

TABLE III. PREDEFINED COLORS FOR COLORS IDENTIFIER (COLORS ARE CAPTURED IN RGB FORMAT THEN CONVERTED TO HSV FORMAT)

Color	Condition Using Hue (H) and Saturation (S) values f(H,S)
White	$H + S > 59$
Black	$H + S < 35$
Red	$H >= 345 or H < 45$
Orange	$45 <= H < 75$
Spring green	$75 <= H < 105$
Green	$105 <= H < 135$
Turquoise	$135 <= H < 165$
Sky Blue	$165 <= H < 195$
Blue	$195 <= H < 225$
Purple	$225 <= H < 285$
Pink	$285 <= H < 315$
Crimson	$315 <= H < 345$



(a) Start Screen. (b) Selecting Device. (c) After Connecting with Device.

Fig. 4. The Interface of the Distance Notifier Application.

Distance notifier application workflow is shown in Fig. 5. After launching the app, users choose the distance identifier device from the list. Upon succession of connection, the app start notifies the user about obstacles in her/his way.

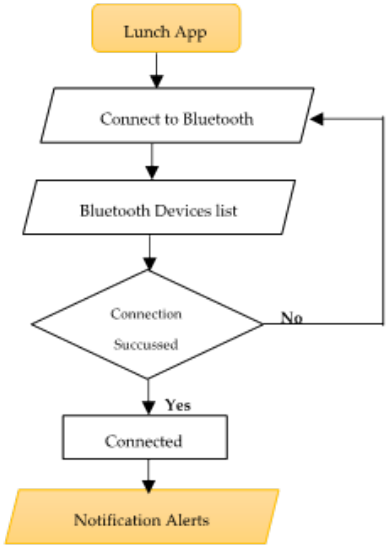


Fig. 5. Flowchart of Distance-Notifier Application.

IV. RESULTS

In this section, we present the evaluation of our experiment which conducted to evaluate the proposed applications' features in the real-world. The design of experiment and reporting of performance results are inspired by [25].

A. Test Cases

To test the proposed applications, we used it in multiple scienaros with various lighting conditions and poses. Our applications have five features: (1) color identifier, (2) object labels, (3) text reader, (4) facial expression, and (5) distance notifier. Fifteen test cases were desgined for each feature to be tested. Fig. 6 shows samples for the test cases with images. Test cases were designed to have veriaty in fabric, color, shape, camera pose and lighting conditions. For each feature 15 test cases was designed.

B. Experimental Results

An experiment is conducted in the real world using the developed two applications. Generally, the application worked with high accuracy as reported in Table IV ranging from as low as 33% and as high as 100%. Reasons vary for this some of which are due very long text, low light, blurred from moving camera or zooming too close or too far in an image. The color identifier is affected by lighting condition a lot as it shows in the accuracy. Future investigation on improving the accuracy of the color identifier is needed.

V. DISCUSSION

There are many ways to improve the accuracy of each feature. Research about each of them is extensive with rapid improvements. However, all this is out of scope of this work.

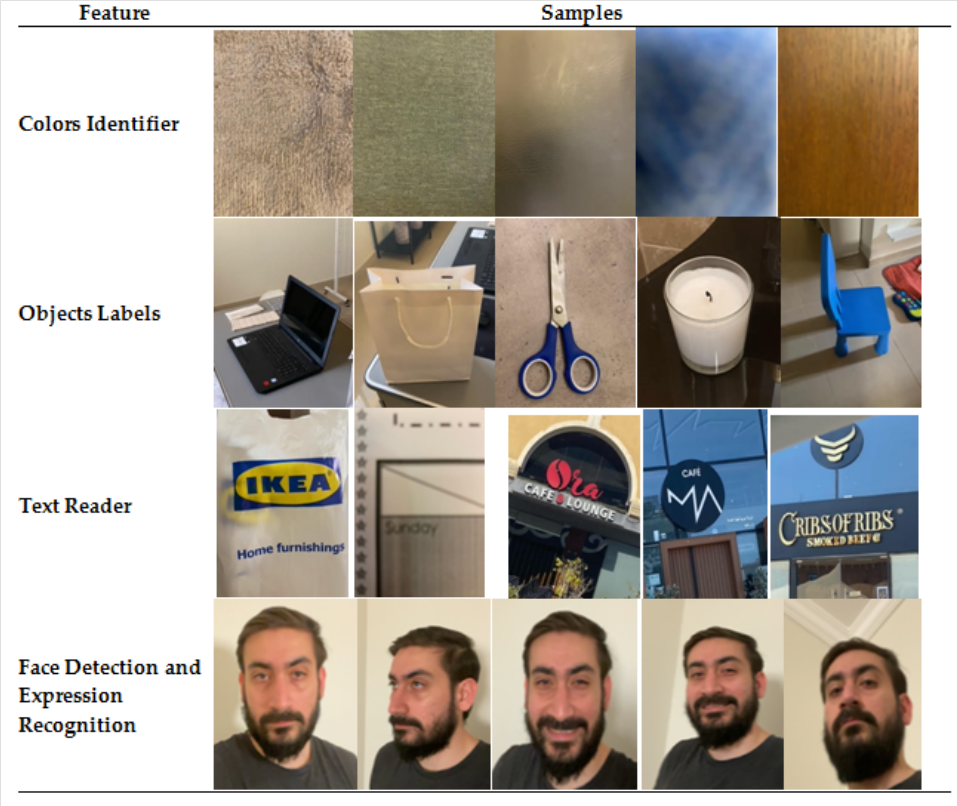


Fig. 6. Test Cases Samples for the Feature Tested.

TABLE IV. EXPERIMENTAL RESULTS FOR ALL TEST CASES FOR EACH FEATURE (INDICATES THAT THE RESULTS ARE REPRESENTED BASED ON THE PERCENTAGE OF LETTERS RECOGNIZE CORRECTLY AS PERCENTAGE OF ALL LETTERS IN THE IMAGE.).

Feature	# Test Cases	Obtained Results	Causes for Errors
Colors Identifier	15	33%	Low light - Blurred image - The image is too close or too far
Objects Labels	15	80%	Low light - Blurred image - The image is too close or too far
Text Reader	15	90%	Very long text - Low light - Blurred image - The image is too close or too far
Face detection	15	100%	
Facial Expression	15	93%	Head pose - Low light - Blurred image - The image is too close or too far
Distance Notifier	N/A	100%	

The focus of this work is to integrate state-of-the-art features that we believe are most beneficial for the application users at low cost. It is also important to note that this application has an interface for interaction without requiring the user to have visual abilities. This expand our application target audience from only partially blind people to complete blind people. One more group of people that could benefits who do not suffer from blindness or short of sight but not color blindness can benefits from this application as well. They can use color identifier when in doubt about the color they see. To this end, our experiment proof the validity of the application and the robustness to various condition in real-world environment.

VI. CONCLUSIONS AND FUTURE WORK

An affordable solution for people with visual impairments was proposed and implemented. Vivid application provides those users the ability to recognize objects, detect people in the scene and their facial expression, assist in identifying colors and help in reading texts. Additionally, an extra feature for obstacle avoidance was implemented using a secondary standalone application with attachable hardware which have relatively low cost. These applications were tested in the real world and provided very good results. The results of the experiments indicate that such an application is a viable option for assisting people in need at an affordable price.

Future work on this research includes improving the use of machine learning to identify colors instead of predefining the range and enhancing the text reader feature to include long texts and more languages for non-English speakers. More work will be done to improve the sensor by shrinking the device size to enhance its portability. Lastly, adding the navigation features to the distance-notifier application which will help the users not only to avoid obstacle, but also to navigate well based on the shorter route. Lastly, more experiments are needed to be conducted with subjects that are from the target user base which can highlight additional challenges and areas of improvements from user experience or robustness of the application.

REFERENCES

[1] R. R. Bourne, S. R. Flaxman, T. Braithwaite, M. V. Cicinelli, A. Das, J. B. Jonas, J. Keeffe, J. H. Kempen, J. Leasher, H. Limburg *et al.*, "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis," *The Lancet Global Health*, vol. 5, no. 9, pp. e888–e897, 2017.

[2] J. Echessa, "A look at android ml kit - the machine learning sdk," https://auth0.com/blog/a-look-at-android-ml-kit-the-machine-learning-sdk/, 2018, [Online; accessed 24-July-2018].

[3] A. S. SINGH D, "Advanced human-smartphone interface for the blind using morse code," *International Journal of Advances in Electronics and Computer Science*, 2017.

[4] R. Kuber, A. Hastings, and M. Tretter, "Determining the accessibility of mobile screen readers for blind users," *UMBC Faculty Collection*, 2020.

[5] GEORGIA, "How to use voiceover for iphone," https://www.imore.com/voiceover-tutorial, 2010, [Online; accessed 17-November-2018].

[6] J. Hildenbrand, "What is google talkback?" https://www.androidcentral.com/what-google-talk-back, 2014, [Online; accessed 11-September-2018].

[7] T. Mayisela, "The potential use of mobile technology: Enhancing accessibility and communication in a blended learning course." *South African Journal of Education*, vol. 33, no. 1, pp. 1–18, 2013.

[8] D. Ahuja, J. Amesar, A. Gurav, S. Sachdev, and V. Zope, "Text extraction and translation from image using ml in android," *International Journal of Innovative Research in Science, Engineering and Technology*, 2018.

[9] S. View, "How do people who are blind or visually impaired read printed text?" https://sandysview1.wordpress.com/2016/09/15/how-do-people-who-are-blind-or-visually-impaired-read-printed-text/, 2016, [Online; accessed 8-September-2018].

[10] A. V. Mhaske and M. S. Sadavarte, "Portable camera based assistive text reading from hand held objects for blind person," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, 2016.

[11] A. K. Joshi, T. P. Madhan, and S. R. Mohan, "Automated electronic pen aiding visually impaired in reading, visualizing and understanding textual contents," in *2011 IEEE INTERNATIONAL CONFERENCE ON ELECTRO/INFORMATION TECHNOLOGY*. IEEE, 2011, pp. 1–6.

[12] M. S. R. Tanveer, M. Hashem, and M. K. Hossain, "Android assistant eyemate for blind and blind tracker," in *2015 18th international conference on computer and information technology (ICCIT)*. IEEE, 2015, pp. 266–271.

[13] W. Elmannai and K. Elleithy, "Sensor-based assistive devices for visually-impaired people: Current status, challenges, and future directions," *Sensors*, vol. 17, no. 3, p. 565, 2017.

[14] C. Willings, "Optical devices for distance," https://www.teachingvisuallyimpaired.com/optical-devices-for-distance.html, 2015, [Online; accessed 12-November-2018].

[15] S. Z. Mohammad Mohammadi, Saif AlAmeri, "Self-energized smart vision stick for visually impaired people," *IEEE Conference on Electromagnetic Field Computation*, 2013.

[16] A. K, "American foundation for the blind," https://www.afb.org/prodProfile.asp?ProdID=746, 2014, [Online; accessed 22-October-2018].

[17] A. Gonzalez, R. Benavente, O. Penacchio, J. Vazquez-Corral, M. Vanrell, and C. A. Parraga, "Study of color importance for blinds, and pda device," *International Journal of Innovative Research in Science*, 2013.

[18] H. Rashid, A. R. Al-Mamun, M. S. R. Robin, M. Ahasan, and S. T. Reza, "Bilingual wearable assistive technology for visually impaired persons," in *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)*. IEEE, 2016, pp. 1–6.

[19] C. Willings, "Identification apps," https://www.teachingvisuallyimpaired.com/identification-apps.html, [Online; accessed 10-October-2018].

[20] "Image labeling features," https://firebase.google.com/docs/ml-kit/label-images, 2018, [Online; accessed 18-October-2018].

[21] K. B. Tharkude, A. K. Wayase, P. S. More, and S. S. Kothey, "Smart android application for blind people based on object detection," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 4, 2016.

[22] P. J. Neha Patil1, "Color to sound converter for blind people," *International Research Journal of Engineering and Technology*, 2017.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[24] Isaac100, "Getting started with the hc-sr04 ultrasonic sensor." https://www.hackster.io/Isaac100/getting-started-with-the-hc-sr04-ultrasonic-sensor-036380, 2017, [Online; accessed 20-September-2018].

[25] G. Senarathne, D. Punchihewa, D. I. Liyanage, G. Wimalaratne, and H. De Silva, "Blindaid-android-based mobile application guide for visually challenged people," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2021, pp. 0039–0045.

Fashion Image Retrieval based on Parallel Branched Attention Network

Sangam Man Buddhacharya[1]
Department of Electronics and Computer Engineering
Institute of Engineering, Pulchowk Campus
Lalitput, Pulchowk, Nepal

Sagar Adhikari[2]
Department of Electronics and Communication Engineering
Paschimanchal Campus
Pokhara, Nepal

Ram Krishna Lamichhane[3]
Department of Electronics and Communication Engineering
Paschimanchal Campus
Pokhara, Nepal

***Abstract*—With the increase in vision-associated applications in e-commerce, image retrieval has become an emerging application in computer vision. Matching the exact user clothes from the database images is challenging due to noisy background, wide variation in orientation and lighting conditions, shape deformations, and the variation in the quality of the images between query and refined shop images. Most existing solutions tend to miss out on either incorporating low-level features or doing it effectively within their networks. Addressing the issue, we propose an attention-based multiscale deep Convolutional Neural Network (CNN) architecture called Parallel Attention ResNet (PAResNet50). It includes other supplementary branches with attention layers to extract low-level discriminative features and uses both high-level and low-level features for the notion of visual similarity. The attention layer focuses on the local discriminative regions and ignores the noisy background. Image retrieval output shows that our approach is robust to different lighting conditions. Experimental results on two public datasets show that our approach effectively locates the important region and significantly improves retrieval accuracy over simple network architectures without attention.**

***Keywords*—Convolutional neural network (CNN); image retrieval; attention mechanism; convolutional block attention module (CBAM)**

I. Introduction

In the last decade, due to our increased computational ability, there have been tremendous improvements in Deep learning [1], [2] and Computer Vision, leading to an exponential proliferation of applicational possibilities. Among the various engineering applications of computer vision ranging from Drug Design [3] to Monocular depth estimation [4], image retrieval has become an emerging one. This particular application has both academic and business ramifications. Academically, it can bring about new innovative approaches to solving image comparison problems, whereas commercially, it can create a disruptive shopping experience for the users. Among all the product categories, due to its dynamic product nature, variations, and immense use case, Clothing/Fashion has received the highest amount of attention.

When similar kinds of images (i.e., consumer to consumer or shop to shop) are compared, there is a certain homogeneity in the images. Thus, they can be treated as from the same domain, not neglecting multiple variations such as lighting, view, backgrounds, product orientation, etc. Nonetheless, comparing different kinds of images (professional with amateur) will contain images from other domains.

Despite the difference in image types, these comparisons can be achieved by analyzing the human-detectable details in the clothes, such as cloth category, color, pattern, prints on the clothes, and so on. Most current retrieval solutions [[5], [6], [7], [8], [9], [10]] incorporate deep learning models that convert actual images into vector representation so that the query image's embedding can be compared against all the images' embeddings from the list, and the closest one can be returned. For that, triplet loss is the most widely used comparative loss technique. As suggested by [[11], [10], [12], [13]], despite being superior to other approaches, the triplet loss approach has its demerits, such as the inability to achieve top performance, being computationally expensive, and being prone to noisy labels and outliers. To mitigate that improvement has been proposed by using the Centroid Triplet Loss function in [14].

Nevertheless, as described in [9], high intra-class variability in clothes and the possibility of different kinds of deformations for the same type of clothes were the significant hurdles for achieving the most acceptable retrieval results. The problem with most of these existing approaches is that it ignores low-level features and those which use low-level features take all the information without selecting discriminative features which introduce noise. Deep networks, which are being used as a solution, tend to go deep and lose vital information from low-level features. Shallow networks can provide those low-level features, but the output is prone to noise. Thus, some form of noise elimination is required. Attention mechanisms emphasize the essential features and suppress the non-essential features. CBAM [15] sequentially applies channel and spatial attention along the respective principle dimensional axes to achieve the same. A shallow network - combined with the attention layer - outputs noiseless low-level features. Thus we have proposed a new architecture that utilizes both deep and attention-shallow networks to extract high-level and discriminative low-level features.

Along with the new architecture proposal, other factors were also considered for improving the overall retrieval accuracy. Here are our contributions:

1) Experimentation with multiple architectures for Image retrieval.
2) Propose a new attention-based architecture for better retrieval performance.
3) Experimentation with the impact of image size on the model's performance.

4) Experimentation with different classification models as our backbone network.
5) Performance comparison across multiple fashion data datasets (DeepFashion [6] and DeepFashion2 [16]).

II. Problem Statement

From a consumer's point of view, there might be different scenarios where a user could benefit from various forms of fashion image comparison and automated searches. All such applications usually include these three kinds of image comparisons:

- Image comparison between a shop image with another shop image.
- Image comparison between a shop image with a consumer image.
- Image comparison between a consumer image with another consumer image.

Due to different image-type comparisons, we prioritize selecting distinctive features in fashion pattern matching, which - moreover - deals with these three main problems in pattern matching:

1) Common images contain different backgrounds, which are usually noisy features for the model. Even after cropping only the target section, the remaining background will still dominate the distinctive features and reduce the model's overall performance.
2) Clothes might contain only a small portion of areas that might cause differentiation from other clothes. Nevertheless, when we use all the features from the clothes to compare the similarity, there might be a low influence of the distinctive features, reducing the performance. Since the distinguishing area varies according to the type of clothes, we need a dynamic module that will focus more on those discriminating features.
3) The existing deep convolution networks - rightfully so - suppress the non-crucial features. While doing so, the low-level features are also being ignored in such a way that it is impacting the retrieval accuracy.

III. Methodology

In this section, we describe our proposed architecture (PAResNet50) with a two branched variation (DBAN) along with loss function, and augmentation policy used during training and testing the network.

A. Architecture

We use a deep Convolutional Neural Network (CNN) to generate feature embeddings. The feature vector is the abstract representation of patterns, color, and shape of the input images, which helps to distinguish between the two different clothes. We use a triplet-based network architecture with the ranking loss function to learn the feature vectors. As shown in the Fig. 1a, the three triplets q, p, and n are independently fed into three different deep CNN, which share similar architecture and parameters. The deep CNN computes respective feature embeddings ($\overrightarrow{q}$, $\overrightarrow{p}$, $and \overrightarrow{n}$) for triples p, q and n.

Inspired by [17], we use multiscale deep CNN. Our implementation is quite different than [17], we use ResNet-50 [18] instead of Alexnet [19] and a series of convolutional and CBAM [15] layers. As shown in Fig. 1b, it has two different parallel branches coupled at conv1 of ResNet-50 [18]. The two parallel branches are downsampled with 4:1 and 8:1 ratios respectively. The downsampled branches are followed by 3x3 convolutional and CBAM [15] layers, flattened to extract low-level features. The output from $conv5\_block3$ of ResNet-50 [18] is followed by a 1x1 convolutional layer and global average layer to extract high-level features. The high-level and low-level features are concatenated and followed by a dense layer to output the final embedding. Introducing an attention mechanism in the shallow branches helps the model to focus on the low-level details like color, texture, and materials regarding its shape. Since the low-level features have lots of noise, reducing the retrieval performance, we used CBAM [15] as the attention module to enhance the essential features while fading out the non-relevant information.

During image retrieval, the embeddings of each image are extracted, and cosine similarity between the embeddings is calculated to find the best matching clothes. Distance between the embeddings estimates the similarity or dissimilarity between the images. Similar images are closer in the embedding space while the dissimilar images are distant.

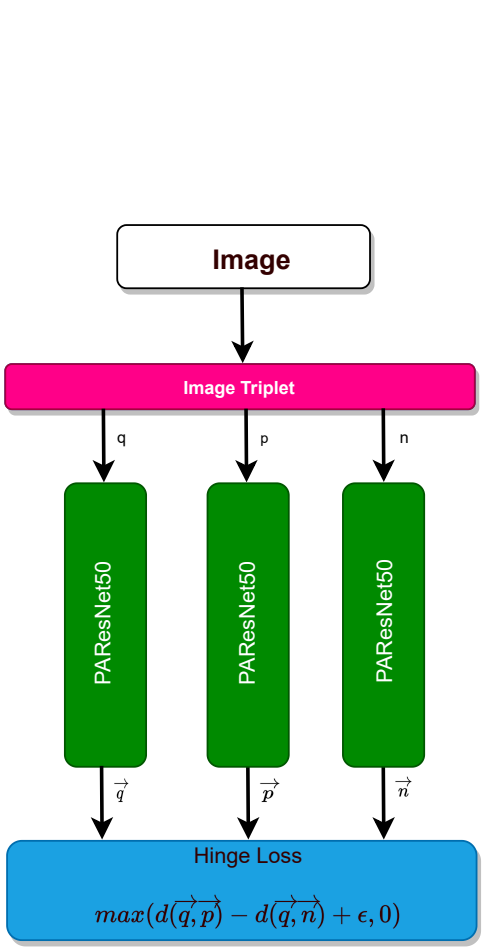
B. Attention Mechanism

Attention mechanism is a technique by which computers try to simulate how human vision focuses in terms of computer-based algorithms. It is a method that tries to enhance the significant parts while fading out the non-relevant information. It can dynamically adjust the weights based on features of the input image.

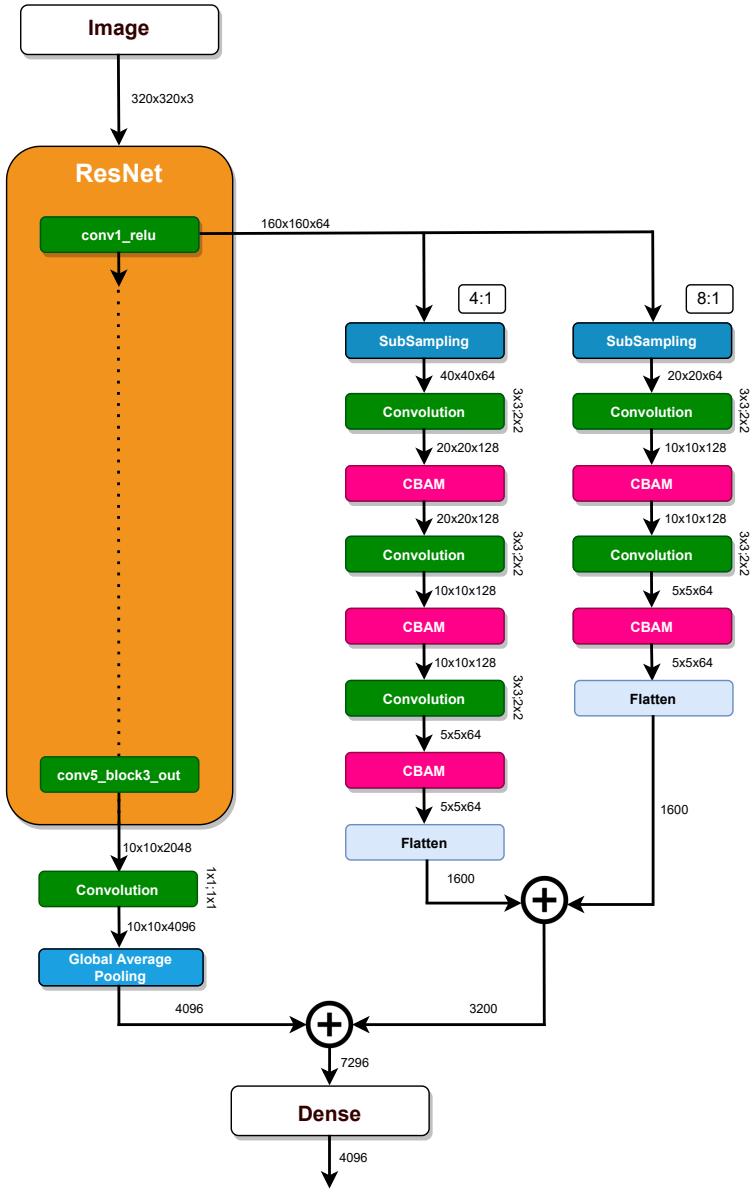
We use Convolutional Block Attention Module (CBAM [15])as our attention module. As in Fig. 2, CBAM [15] is composed of two sequential sub-modules, the Channel Attention Module (CAM) and the Spatial Attention Module (SAM).

a) Channel Attention Module: Channels are feature maps stacked in a tensor, where each cross-sectional slice is, basically, a feature map of dimension (h x w). The input feature map of the channel is regarded as a feature detector. Channel attention is calculated by compressing the feature map in the spatial dimension using max pooling and average pooling to obtain two different spatial context descriptors. The descriptors are fed into a shared network to produce a feature vector. The shared network comprises an MLP (Multi-Layer Perceptron) and one hidden layer. The output feature vectors from MLP are merged using element-wise summation, and the sigmoid function is applied to compute the channel attention map.

b) Spatial Attention Module: Spatial attention represents the attention mechanism masks on a single cross-sectional slice of the tensor or each feature map representing the Spatial Attention Map. As in Fig. 2, Spatial attention is calculated with the two different feature descriptions obtained from maximum pooling and average pooling in the channel dimension. The two feature descriptions are merged, and a convolutional operation is applied to generate a spatial attention map.



(a) Deep Ranking Architecture



(b) Dual Branched Attention Network of PAResNet50

Fig. 1. Overall Architecture of PAResNet50 with Deep Ranking

C. Loss Function

We have used a triplet loss function with batch-all online-mining strategies. A batch of B embedding is extracted from a batch of B inputs. B is composed of C different styled clothes with N images each. A valid triplet $\overrightarrow{q}, \overrightarrow{p}, \overrightarrow{n}$ is generated from B embeddings. These three indices $(\overrightarrow{q}, \overrightarrow{p}, \overrightarrow{n}) \in [1, B]$ are query, positive and negative pairs, respectively. Batch all online mining produces a total of T (1) valid triplets

$$T = C * N * (N-1) * (C * N - N) \quad (1)$$

where $C * N$ is the number of query images, $N-1$ is the possible positive pair per query images and $C * N - N$ is the possible negative pair. Hinge loss is calculated from each valid triplets $(q, p, n) \in [1, B]$.

$$l(\overrightarrow{q}, \overrightarrow{p}\,\overrightarrow{n}) = max(d(\overrightarrow{q}, \overrightarrow{p}) - d(\overrightarrow{q}, \overrightarrow{n}) + \epsilon, 0) \quad (2)$$

where, ϵ is the margin and $d(\overrightarrow{x}, \overrightarrow{y})$ is the Euclidean Distance between $\overrightarrow{x} and \overrightarrow{y}$. The hinge loss function tries to push $d(\overrightarrow{q}, \overrightarrow{p})$ to 0 (i.e. pulling $\overrightarrow{q} and \overrightarrow{p}$ closer) and $d(q, n)$ to be greater than $d(\overrightarrow{q}, \overrightarrow{p}) + \epsilon$ (i.e. pushing $\overrightarrow{q} and \overrightarrow{n}$ farther). Our final training loss L is as follows:

$$L = \sum_{(\overrightarrow{q}, \overrightarrow{p}\,\overrightarrow{n}) \in B}^{T} l(\overrightarrow{q}, \overrightarrow{p}\,\overrightarrow{n}) \quad (3)$$

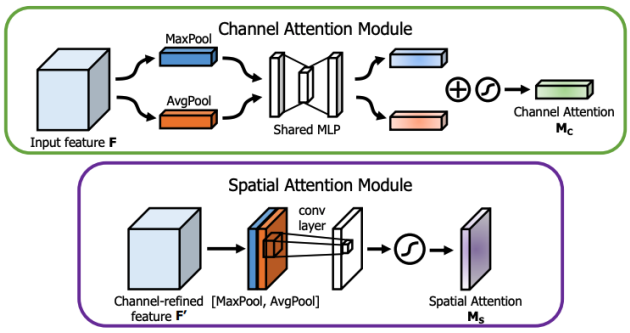


Fig. 2. Structure of Spatial and Channel Attention in CBAM. Source: [15].

D. Training Data Generation

For both DeepFashion [6] and DeepFashion2 [16] datasets we used the provided benchmark training sets and placed clothes with same style in a common style folder. These style folders were kept in their respective category folder. Each image was cropped with provided bounding boxes. For DeepFashion2 [16], back-faced and heavily occluded data were removed. We created a list of all the images available in the folder. Two images from each style/group were randomly selected during training to create a batch. The selected pairs were excluded from the list until the next epoch. Epoch is completed when there is no image pair left in the list.
We have only used geometric augmentation for both the query and shop images. The input images are horizontally flipped with 50% chance and rotated randomly in a range of $[-1, 1]$ degrees. This helps to increase the generalization performance and avoid over-fitting. Colour augmentation might change the original color of both query and shop images which might cause the corresponding pairs to be dissimilar, so we didn't use colour augmentation.

IV. Experiments

A. Datasets

a) DeepFashion [6]: The dataset contains over $800,000$ images with the information of categories, landmarks, bounding boxes, clothes attributes, and image pairs for Consumer-to-Shop/In-shop clothes retrieval. For this paper, we have only used the Consumer-to-Shop Clothes Retrieval subset which contains $33,881$ unique clothing products, $239,557$ consumer and shop images and $195,540$ consumer and shop matching pairs.

b) DeepFashion2 [16]: The dataset contains $491k$ diverse images from both consumers and shopping where each item is labeled with scale, occlusion, zoom-in, viewpoint, category, style bounding box, dense landmark, and per-pixel mask. For this paper, we only use Commercial-Consumer clothes pairs which continents $319k$ training sets, $34k$ validation sets, and $67k$ test sets. From the available dataset, we removed back-face and heavily occluded clothes during training.

B. Implementation Details

For the implementation, Keras [20] and TensorFlow [21] have been used as our deep learning framework. Likewise, Adam optimizer has been used to train the model with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-6}$ and an exponential decaying learning rate of 0.96 for every 150000 steps with starting learning rate of 10^{-6}. For online triplet loss, margin of 1.5 has been set. We have used a batch size of 4 composed of 2 different styles of clothes with 2 images each. Each model has been trained for different iterations; the training is stopped according to the model's performance on validation loss. All the experiments have been performed in Kaggle with NVidia K80 GPU.

C. Evaluation Metrics

For the evaluation of retrieval performance, we use top-k accuracy, as in [[22], [6]]. The top-k accuracy is defined as follows:

$$P(K) = \frac{\sum_{q \in N} hit(q, K)}{|N|} \tag{4}$$

where, N is the total number of queries performed.

$hit(q, K) = 1$ is a hit, if at least one shop image appears within the top-K ranking for the query image q.

$hit(q, K) = 0$ is a miss, if no any shop image appears with in the top-K ranking for the query image q.

D. Experiments with Different Embedding Layers

In this experiment, we have used different embedding layers keeping other parameters unchanged. We used Flatten layer, Spatial Pyramid Pooling(SPP) layer, and Global Average layer after $conv5\_block3$ of ResNet-50 [18]. Table I shows that the flatten layer has the highest number of feature vectors with the lowest accuracy. But the global average layer has less number of feature vectors with the highest accuracy. In the flatten layer, redundant features and noise reduced the influence of discriminative features. But in the global average layer, there are mostly discriminative features. Therefore the retrieval performance depends upon the size of the feature vector. We didn't find SPP efficient compared to Global Average, so we used GlobalAverage as our embedding layer to extract high-level features.

TABLE I. Comparison of Top-k (k= 1, 5, 10, 20, 50) Retrieval Accuracy on DeepFashion2 [16] Dataset for Different Embedding Layers Performed on 256x256 Image Size.

Last layer	# size	mAP	top-1	top-5	top-10	top-20	top-50
Flatten	65536	0.687	0.445	0.629	0.712	0.784	0.865
SPP	21504	0.720	0.485	0.663	0.743	0.815	0.893
GlobalAvg	4096	**0.785**	**0.576**	**0.747**	**0.812**	**0.863**	**0.927**

E. Experiments with Different Backbone Networks

In this section, we have experimented with different classification models to find the best retrieval performance. From Table II, it can be clearly observed that ResNet-50 [18] architecture has significantly higher performance in comparision to VGG-16 [23], and MobileNetV1 [24], so we used ResNet-50 [18] as our backbone network in PAResNet50 [1].

TABLE II. COMPARISON OF TOP-K (K= 1, 5, 10, 20, 50) RETRIEVAL ACCURACY ON DEEPFASHION2 DATASET FOR DIFFERENT ARCHITECTURES PERFORMED ON 256X256 IMAGE SIZE.

Models	mAP	top-1	top-5	top-10	top-20	top-50
VGG-16 [23]	0.699	0.453	0.633	0.715	0.804	0.894
MobilenetV1 [24]	0.566	0.315	0.486	0.572	0.665	0.793
ResNet-50 [18]	**0.798**	**0.588**	**0.761**	**0.822**	**0.882**	**0.937**

TABLE III. COMPARISON OF TOP-K(K=1,5,10,20,50) RETRIEVAL ACCURACY ON DEEPFASHION2 DATASET FOR DIFFERENT IMAGE SIZES.

Image size	mAP	top-1	top-5	top-10	top-20	top-50
256x128	0.7896	0.567	0.746	0.817	0.879	0.939
256x256	0.798	0.588	0.761	0.822	0.882	0.937
320x320	**0.813**	**0.617**	**0.774**	**0.834**	**0.895**	**0.943**

F. Experiments with Different Image Size

To find the influence of image size in PAResNet50, we have experimented with different image sizes while keeping other parameters constant. From Table III, we found the input images of size 320x320 to be the best for our settings. Therefore, a larger image size helps to increase the retrieval performance so we used 320x320 image size in PAResNet50 for both training and testing.

G. Experiments with Different Architectures

We experimented with different kinds of architectures. They are as follows:

a) Simple Network(SN): It is a simple ResNet-50 [18] classification model pre-trained on Imagenet [25]. The output from $conv5\_block3$ of ResNet-50 [18] is followed by 1x1 convolutional layer, global average layer and a dense layer to extract a feature embeddings.

Fig. 3. a) and b) are the Top-1 and Top-5 Categories Retrieval Accuracy on DeepFashion [6] Validation Set. Each Model is Trained on Image Size of 320x320.

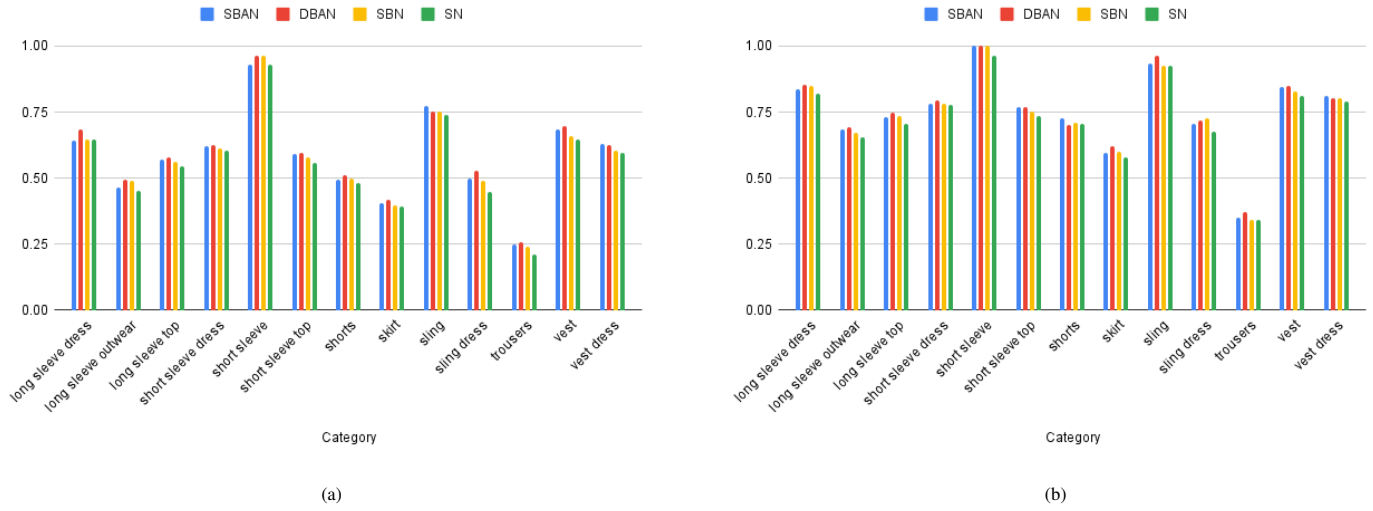
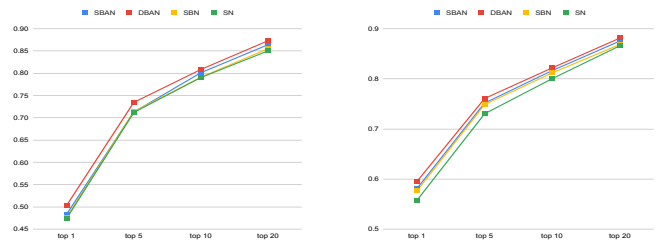


Fig. 4. a) and b) are the Top-1 and Top-5 Categories Retrieval Accuracy on DeepFashion2 [16] Validation Set. Each of the Model is Trained on Image Size of 256x256.

b) Single Branched Network(SBN): It is an extention to already existing Simple Network. A ResNet-50 [18] classification model coupled with a parallel branch. In the parallel branch, the output from $conv1\_relu$ is downsampled with a ratio of 4:1 and a series of convolutional layers is used. The output from the global average layer and the parallel branch is concatenated which is followed by a dense layer to extract the final feature embeddings.

c) Single Branched Attention Network (SBAN): This follows the architecture of Single branched network (SBN) here the convolutional layer in the parallel branch is followed by the CBAM [15] layer.

d) Dual Branched Attention Network (DBAN/PAResnet50): It is our final model, which has shown the best performance. It has two parallel branches with downsampling of 4:1 and 8:1, respectively. After downsampling on each branch, a series of convolutional and CBAM [15] layer is used which is followed by a flatten layer to extract low-level features. The outputs from the global average layer and the two parallel branches are concatenated and followed by a dense layer to extract the final feature embeddings. The architecture of PAResNet50 is shown in the Fig. 1.



(a) (b)

Fig. 5. a) and b) are the Comparison of Top-k(k=1,5,10,20) Retrieval Accuracy for Different Architecture on DeepFashion [6] and DeepFashion2 [16] Dataset Respectively. Each of the Model is Trained on Image Size of 256x256.

e) Architecture Comparison: From the Fig. 5, we observed that using low-level features directly from the branched network (SBN) slightly increases the model's performance compared to Simple Network (SN). The increase in performance is due to addition of low-level features. Adding a CBAM [15] layer in the branch ($SBAN$) further improves the performance since the attention mechanism suppresses the noises from low-level features. When two branches with an CBAM [15] layer ($PAResNet50$) are used to extract the low-level features, the model gets on additional fashion details (i.e., color, styles, and patterns) to learn, which significantly increases the retrieval performance. Therefore Dual Branched Attention Network(DBAN/PAResnet50) has higher retrieval accuracy in comparison to other architectures. We also applied attention mechanism on high level features by adding CBAM [15] layer on different blocks of ResNet-50 [18], but it didn't improve the performance. The attention mechanism didn't work well on high-level features.

To better analyze each architecture's performance, we evaluated the top-1 and top-5 retrieval accuracy for each category on both Deepfashion [6] and Deepfashion2 [16] datasets. Fig. 3 shows that PAResnet50 has improved the top-1 and top-5 retrieval accuracy on DeepFashion [6] for clothing, dresses, and tops while slightly improving in trousers. From the Fig. 4, we can see that on DeepFashion2 [16], PAResnet50 has the highest top-1 retrieval accuracy in all categories except sling and vest dress. In the top-5 retrieval accuracy, it has also performed well in the sling category.

H. Results on Deepfashion [6] and Deepfashion2 [16] Dataset

TABLE IV. COMPARISON OF PARESNET50'S [1] TOP-K (K=1,5,10,20,50) RETRIEVAL ACCURACY ON DEEPFASHION [6] AND DEEPFASHION2 [16] DATASETS WITH 320X320 IMAGE SIZE

Datasets	mAP	top-1	top-5	top-10	top-20	top-50
DeepFashion [6]	0.771	0.503	0.733	0.810	0.873	0.936
DeepFashion2 [16]	0.813	0.617	0.774	0.834	0.895	0.943

We trained PAResNet50 on both DeepFashion [6] and DeepFashion2 [16] datasets with image size of 320x320. Table IV shows that DeepFashion2 [16] dataset has higher performance in comparison to DeepFashion [6] since we have removed the back-faced and highly occluded images in DeepFashion2 [16], which reduced the conflict invalid image pairs. The back-faced and occluded clothes might have different colors, patterns, and texture, so when paired together, it forms invalid pair and decreases the training performance.

Results from Table IV confirm that our proposed model PAResNet50 is suitable for fashion image retrieval on different e-commerce websites.

I. Query Results

To better understand the output quality of PAResNet50, we analyzed the query results on different category images as shown in Fig. 6. The output is categorized into three groups best, good, and bad. The top three rows are the best output, retrieving the corresponding shop image in the top-1 list. The fourth and fifth rows are the good outputs, retrieving the corresponding shop image in the top-3 list. The bottom row is the bad output where the pair shop doesn't occur within the top-3 list. We can observe that our model can retrieve perfect matching images by learning fashion details such as colors, styles, patterns, and textures. In the second row first query, our model has retrieved the exact shop image even if the cloth is not worn (shape deformed). With results from the first-row second query and second-row second query, we can see that even under different lighting conditions, our model has delivered the exact shop image in the top-1 list. Therefore, our model is robust to different lighting conditions. In the second last row of Fig. 6, although the exact shop image is not retrieved in the top-1 list, visually similar colors and pattern-styled clothes are retrieved, which is a more challenging task for a human being. In the bottom row, even though the exact shop image doesn't appear in the top-3 list, the retrieved images are significantly similar to the query image.

J. Attention Visualization

To find the effect of the attention mechanism (CBAM [15]), we have visualized the attention map from PAResNet50. We

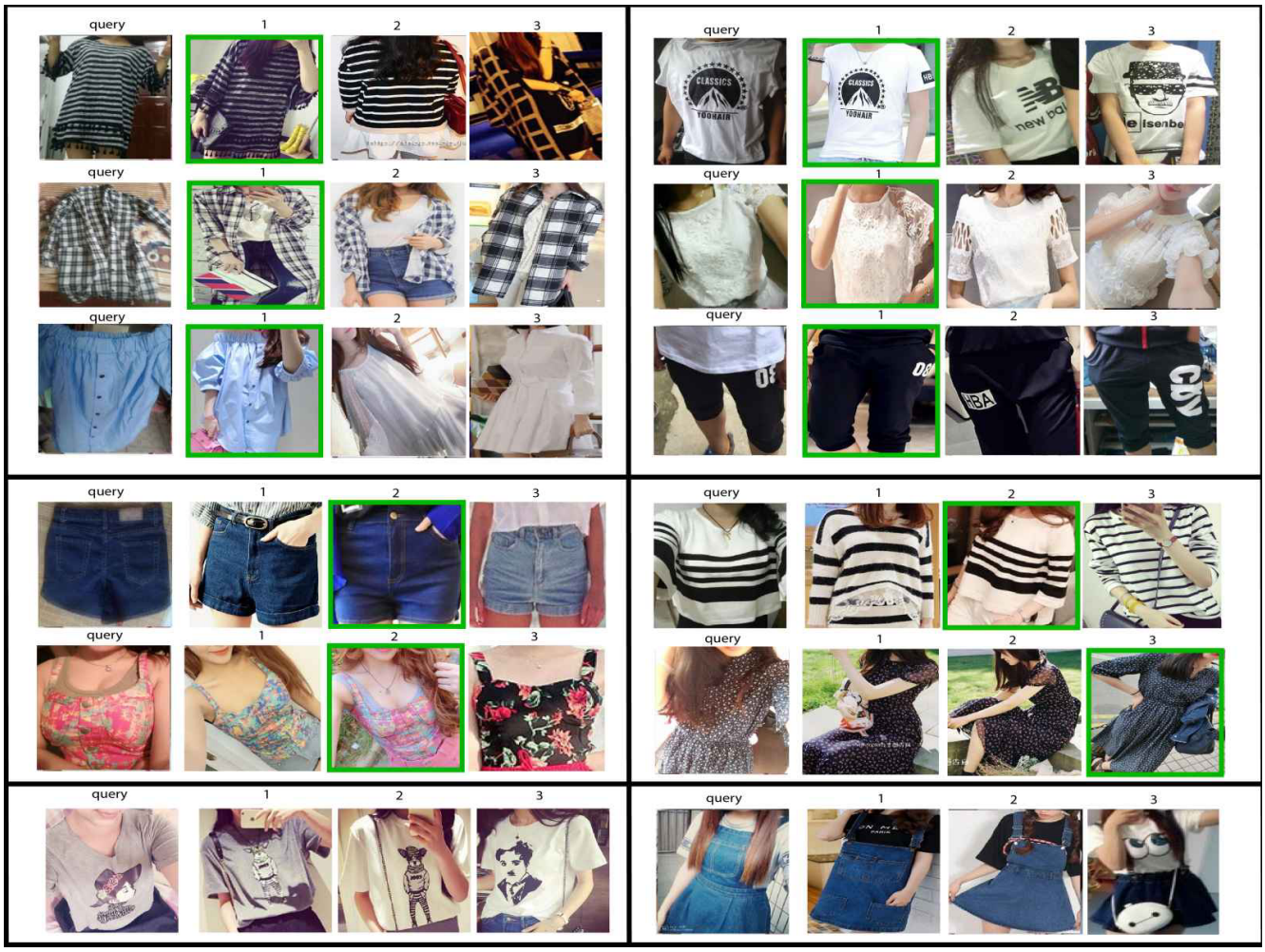


Fig. 6. Top-3 Retrieved Images for a Given Query on DeepFashion2 [16] Dataset. Green Box Indicates the Corresponding Shop Image.

Fig. 7. Visualization of Attention Map in the Query Images from PAResNet50 1. Red Indicates Higher Important Region while Blue Indicates Lower Important Regions.

used Grad-CAM [26] for visualization. From Fig. 7, we can clearly observe that our model has mainly focused on the local discriminative regions (e.g. logos, pictures, patterns, and text) in an input image while ignoring non-discriminative regions (e.g. background, plain region, and hand). Therefore only the discriminative features are used to find the matching images, which increases the retrieval performance of the model. Attention mechanism on the branch layer helps the network focus on only the important features while ignoring the less significant ones.

K. Experimental Summary

Overall, multiple experiments were conducted to find the best settings for image retrieval tasks. Table II, which is the comparison of different classification models (VGG-16, MobileNet, and ResNet-50), shows that ResNet-50 outperforms other models with a minimum margin of 10 percent in mAP metrics. Likewise, Table I clearly depicts that using ResNet-50 architecture with Global Average as embedding layer has performed the best with top-k (k=1, 5, 10, 20, 50) accuracy as 0.576, 0.747, 0.812, 0.863, 0.927 respectively. Further, to show the importance of low-level features and attention mechanisms in image retrieval tasks, we experimented with different architectures. Experimental results from Fig. 5, clearly indicate that Dual Branched Attention Network (DBAN) has achieved the highest retrieval accuracy. Analyzing the Fig. 3 and Fig. 4 demonstrates that DBAN works best in almost all categories. Also, the experiment concluded to observe the influence of different image sizes displays that higher resolution increases the model retrieval performance. As shown in Table III, an image size of 320x320 works best for DBAN. The query output from Fig. 6 helps to better understand the quality of DBAN which shows that this model retrieves visually similar colors and pattern-styled clothes and is robust to different lighting conditions. To further show the attention region of the DBAN, we have visualized the attention map in Fig. 7. We observed that the model has primarily focused on discriminative features. Therefore, it confirms that the attention mechanism ignores the noisy background.

V. Conclusion

In this paper, we have designed the PAResNet50 architecture to present the importance of the low-level features with an attention mechanism for image retrieval tasks. We found that two coupled attention branches in Dual Branched Attention Network(DBAN/PAResNet50) learn low-level fine details and effectively locate the local discriminative regions while ignoring non-significant areas. From various experiments, it can be inferred that incorporating low-level discriminative features along with high-level features improves retrieval performance. The query results exhibit the usability of PAResNet50 in a variety of categories for different e-commerce purposes. Experiments with different architectures(SN, SBN, SBAN, and DBAN) on two public datasets, DeepFashion, and DeeFashion2, demonstrate that DBAN(PAResNet50) outperforms other architectures with fewer or no attention branches. This result leaves room for the possibility of future enhancement in the retrieval accuracy by experimenting with a greater number of such multiscale attention branches.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] K. Aryal, M. Gupta, and M. Abdelsalam, "A survey on adversarial attacks for malware analysis," *arXiv preprint arXiv:2111.08223*, 2021.

[3] A. Dhakal, C. McKay, J. J. Tanner, and J. Cheng, "Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions," *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab476, 2022.

[4] S. M. Buddhacharya, R. Adhikari, and N. Maharjan, "Monocular depth estimation using a multi-grid attention-based model," *Journal of Innovative Image Processing*, vol. 4, no. 3, pp. 127–146, Aug. 2022. [Online]. Available: https://doi.org/10.36548/jiip.2022.3.001

[5] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1218–1226.

[6] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1096–1104.

[7] H. Jun, B. Ko, Y. Kim, I. Kim, and J. Kim, "Combination of multiple global descriptors for image retrieval," *arXiv preprint arXiv:1903.10663*, 2019.

[8] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2019.

[9] M. Wieczorek, A. Michalowski, A. Wroblewska, and J. Dabrowski, "A strong baseline for fashion retrieval with person re-identification models," in *International Conference on Neural Information Processing*. Springer, 2020, pp. 294–301.

[10] Y. Yuan, W. Chen, Y. Yang, and Z. Wang, "In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 354–355.

[11] T.-T. Do, T. Tran, I. Reid, V. Kumar, T. Hoang, and G. Carneiro, "A theoretically sound upper bound on the triplet loss for improving the efficiency of deep distance metric learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 404–10 413.

[12] Z. Zhang, C. Lan, W. Zeng, Z. Chen, and S.-F. Chang, "Rethinking classification loss designs for person re-identification with a unified view," *ArXiv abs/2006.04991*, 2020.

[13] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3741–3750.

[14] M. Wieczorek, B. Rychalska, and J. Dabrowski, "On the unreasonable effectiveness of centroids in image retrieval," in *International Conference on Neural Information Processing*. Springer, 2021, pp. 212–223.

[15] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 3–19.

[16] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5332–5340.

[17] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[20] F. Chollet *et al.* (2015) Keras. [Online]. Available: https://github.com/fchollet/keras

[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[22] J. Huang, R. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1062–1070.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[24] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *ArXiv*, vol. abs/1704.04861, 2017.

[25] 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.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

Watchdog Monitoring for Detecting and Handling of Control Flow Hijack on RISC-V-based Binaries

Toyosi Oyinloye[1], Lee Speakman[2], Thaddeus Eze[3] and Lucas O'Mahony[4]
Department of Computer Science
University of Chester, Chester, UK[1,3,4]
School of Science, Engineering and Environment
University of Salford, Manchester, UK[2]

***Abstract*—Control flow hijacking has been a major challenge in software security. Several means of protections have been developed but insecurities persist. This is because existing protections have sometimes been circumvented while some resilient protections do not cover all applications. Studies have revealed that a holistic way of tackling software insecurity could involve watchdog monitoring and detection via Control Flow Integrity (CFI). The CFI concept has shown a good measure of reliability to mitigate control flow hijacking. However, sophisticated attack techniques in the form of Return Oriented Programming (ROP) have persisted. A flexible protection is desirable, which not only covers as many architecture structures as possible but also mitigates known resilient attacks like ROP. The solution proffered here is a hybrid of CFI and watchdog timing via inter-process signaling (IP-CFI). It is a software-based protection that involves recompilation of the target program. The implementation here is on vulnerable RISC-V-based process but is flexible and could be adapted on other architectures. We present a proof of concept in IP-CFI which when applied to a vulnerable program, ROP is mitigated. The target program incurs a run-time overhead of 1.5%. The code is available.**

***Keywords*—Watchdog; return oriented programming; RISC-V; control flow integrity; software security**

I. Introduction

Securing software from hijacking and exploitation is a major step in software development lifecycle and has been faced with persistent challenge especially in the area of control flow hijacking. Attacks via Return Oriented Programming (ROP) [1] remain a source of concern in spite of existing basic and sophisticated protections. Basic protections like DEP/NX [2], which are generic do not mitigate ROP. This is because they are centered around blocking execution of injected code. Although these provide some measure of protections, attacks that stem from code reuse [3] like ROP which are not detectable by memory protection mechanisms, cannot be stopped by data execution prevention. On the other hand, the original CFI [4] relies on the precision of the Control Flow Graph (CFG). The CFG facilitates the detection of abnormal behaviour in the protected process but [4] inaccuracies in the CFG is one of the limitations cited by [4]. Besides, the classic CFI is a non-generic solution. Another non-generic CFI-based solution is Modular Control Flow Integrity (MCFI) [5] which offers a reasonably high level of precision. Recent studies also proffered solution in the form of PUFCanary FIXER [6], a hardware-based CFI which is also limited due to possible information leak where the PUFCanary FIXER inherits known canary limitations. There are other variations of CFI implementation which have contributed positively to the efforts to combat Control Flow Hijacks (CFH) but limitations exist. This could be due to specificity in the architecture that the solution was built on, as we have it in [4], or variation in source code language of target program as we have in MCFI [5], or general cost of implementation for hardware reliant fixes as exist in [7, 8], or the overhead incurred. Gaps and limitations in the realistic adoption of existing solutions inspire the continuous search for adaptable protection for vulnerable applications against attacks like ROP.

Aside from these limitations in existing protection techniques, studies on software protection are mainly implemented on specific system architectures focusing on elements that are involved in the execution of applications. Previous studies on protection measures have mostly focused on earlier architectures like x86 [4] and ARM [9]. This is justifiable because in past years, attention has been given to securing computers and servers which are mostly built on x86. However, in recent years, new technologies have emerged which require more options of protections. The RISC-V [10] technology is one of such which in recent years has gained popularity among producers of CPUs for automotive, smart devices, health tracking devices, etc., [11] because it is open source and more affordable. Also recently, the first laptop running on RISC-V processor has been introduced [12]. With these advancements in technology, a proactive measure of protection is desirable for vulnerable applications and the infrastructures on which they reside. RISC-V is an open-sourced instruction set architecture (ISA) and it was built on the already-established RISC technology [13]. Unlike most other ISAs, RISC-V was designed by academics and was made to be flexible and affordable, not only for use in academic research but also for deployability in hardware and software designs without incurring any royalties. For this reason, producers of embedded device, smart devices, etc., have opted for it. Not much attention has been given to securing RISC-V compared to other architectures like x86. Existing protections might not adequately provide the needed protection for RISC-V- based programs and systems.

Recent studies [14, 15] have highlighted gaps in existing protections especially for RISC-V programs and specifically against ROP as a result of hidden execution paths in the Control Flow Graph applied for implementing CFI. This study was embarked on with the goal to fill the gaps by increasing adaptability of protection mechanism for surmounting ROP. The concept implemented in this paper builds on a previous study [16] which proposed the possibility of securing vul-

nerable processes via inter-process communication. A novel approach was derived as Inter-process CFI (IP-CFI) which adopts the CFI concept alongside inter-process signaling and watchdog monitoring to detect abnormal behaviour in vulnerable applications. The vulnerable process is monitored by another process during execution. If a deviation is suspected in the control flow of the process, a watchdog function and inter-process signaling is triggered to further handle control flow monitoring.

According to [16], in-line CFI could be implemented by inserting labels to mark the start and end of each function with some additional instructions to perform checks on the flow of execution. In building on this technique, the watchdog adopts the time-out concept to extend monitoring whenever the process exceeds a stipulated time frame. The idea of watchdog monitoring is not new for securing systems. For example, study by [17] presents the *grenade* for monitoring mobile apps especially against denial of service (DoS) attacks. The research [17]'s *grenade* uses a countdown timer which is not reloadable once it begins to countdown. The author in [17] opted for this same technique to avoid a hijack of processors where a malicious program is able to extend its own life. In a similar line of thought, we avoid a possible extension of any malicious code execution by running a waiting time based on the intended purpose of the protected process. This is because unlike *grenade* which relies on the operating system timer, IP-CFI uses a monitoring program that is dedicated for monitoring a target process to increase flexibility. Some waiting time is also triggered in the target process and the monitor to achieve inter-process signaling. If a CFH is detected, further exploits can be prevented by an outright halting of the process.

The detection is made possible by initially analysing the vulnerable program to chart its intended execution path representing the CFG of the program, through static and dynamic analysis. In IP-CFI, values are passed from in-line CFI into shared memory where the monitor performs status check of the vulnerable process. Inter-process communication is achieved using atomic operation via semaphore and mutex on shared memory. Values that are used in the monitoring processes are stored in immutable registers and set in assembly code before completing compilation. Since IP-CFI is a software-based implementation which involves addition of enhancement code, the target program would require rebuilding after appending the enhancement code to implement the new protection.

This paper discusses the use of static analysis, dynamic analysis, RISC-V assembly coding, insertion of in-line checks for IP-CFI which provides a behaviour-based detection, and handling of CFH via Inter-process signaling on programs built and run on the RISC-V architecture. Most similar existing solutions are centered on the x86 system architecture and lack capacity to protect applications running on other CPUs like RISC-V. For this reason, the solution presented here is built around the RISC-V architecture but could be adapted for programs running on other CPUs. The rest of this paper is structured beginning with related works discussed in Section II. The methodology and implementation details are held in Section III while Section IV highlights the outcome of implementation, evaluation and application. Section V is a conclusion on this study and possible future works.

II. Related Works

Studies on software exploitation and protection have revealed control flow hijacks as a major source of concern in software security. Researchers have identified strengths and weaknesses of existing mitigation techniques by demonstrating various instances and concepts. The issue of CFH is particularly complex because there are different factors that need to be considered in proffering a lasting solution. This could be the consideration of the programming language of build, the low precision of CFG, CPU architecture on which the applications are running, and the cost of applying hardware solutions. The protection offered through CFI has potentials if applied alongside external enhancements. According to [18], CFI being a concept is flexible and could be enhanced by additional operations.

The classic CFI which was implemented on x86 architecture by [4] presented a promising solution to the challenge of CFH. A concept that relies on expected behavior, detection of deviations from expected behavior and trustworthiness of detector/enforcer. The classic CFI concept makes use of CFGs to apply inline reference monitoring (IRM) with which the protected application is rewritten. This was however found to be inefficient in its fine-grained form and not realistically implementable. Since the outcome of study by [4], other implementations have been studied in [5, 19, 6, 8], etc. These held some reliable outcomes with variations in structure, model, and platform but the mechanism still involves cross-checking flow of execution in comparison to intended flow. The integrity of the process is then enforced by introducing a halt or other forms of handlers to the situation.

Aside from existing limitations is realistic implementation of the classic CFI, recent studies [14, 15] have revealed the possibility of Hidden Execution Paths (HEP) which are not detectable and therefore omitted in the mapping when a CFG is built. While addressing ROP as a threat model on RISC-V, [14, 15] identified how ROP could persist on RISC-V platforms as a result of gadgets that could be obtained from overlapping code. It is therefore desirable to have a protection that is capable of an overview of the protected program. IP-CFI does not seek to know what gadgets are involved in the attack but to ensure the continuity of genuine execution and the termination of illegitimate flow in execution. Previous forms of CFI [4, 20] have their protection mechanism lying within the protected binaries which to an extent provides an impactful protection but the CFI itself might be unreliable due to low precision in CFG. In the case of [20], HEPs that were recently identified [14] could enable attack bypassing the checks. In this study, an additional monitoring process is adopted so that the in-line CFI could be monitored from outside of the target process while a watchdog is triggered if a deviation is suspected.

We present here a software-based protection. The classic CFI [4] was also software-based and was accomplished without recompilation of the program and no access to source code. This was made possible on the x86 implementation because CFG that was used in performing CFI checks were built using Vulcan [21]. Vulcan is not yet compatible with the RISC-V environment and as the program used here is a simple one, the CFG was done using Ghidra and Gdb for analysis. On the other hand, there exists compiler-based implementations like Gfree [20] which requires a part of the binary to be rewritten

and recompiled because the solution aims to eliminate gadgets that are based in libraries. Similarly, the protection here applies some additional lines of instruction to the protected target at the assembly level and also requires a full compilation into executable after the enhancement code has been inserted. The insertion of code has been automated but in-line checks are still inserted manually. A RISC-V target C-source code can be passed as argument into a startup script to be compiled with the enhancement code. The monitor however runs separately and need to be run concurrently as with the target program.

A state-of-the-art study presented in [18] observed that all eleven software-based CFI that were examined could be bypassed, although they each provided some protections in one way or another. They identified fine-grained CFI as a strong defense but incurs high overhead because of the use of shadow stack. Coarse-grained CFI on the other hand is a looser form that checks if control-flow transfers has originated from a return instruction and if its destination can be targeted. Three hardware-based CFI were examined and they were identified as difficult to be realistically implementable as such approach requires changes to the IT ecosystem that would incur additional cost on the system. [18] made conclusions that a hybrid form of CFI that combines existing protections might improve security in a CFI protected program. This study aims to adopt this suggestion by combining the use of In-line CFI, Inter-Process Monitoring (IPM), and watchdog time-out.

Another recent study [6] presents FIXER for protecting RISC-V applications using hardware. Subsequent improvement on FIXER involves the use of a PUFCanary [22]. This however had its own limitations in that a diversion may occur before the canary check, also FIXER does not protect against memory disclosure and it may cause the custom instruction to be bypassed.

[17] came up with a study back in 2000 with foresight on the advancement in technology in the future which we are now in. They foresaw a future where the use of computerised devices would become the norm and based on this, they presented the idea of *grenade* based on the concept of watchdog timer to protect against malicious mobile apps and ensure stability in services running on vulnerable systems. The use of a watchdog timer is a reasonable option for combatting a variety of attacks including Denial of Service (DoS) attacks. This is a relatable scenario as we find that ROP attacks on RISC-V, when chained in some particular order could lead to denial of service. The protection presented here adopts a similar approach by applying the watchdog concept alongside in-line CFI to ensure that such DoS attacks are detected and handled.

During this study, we identify RISC-V ROP gadgets that cause the denial of service. Among numerous possibilities of the outcome of ROP, DoS could occur when chained ROP gadgets don't include an instruction to redirect execution to a location where other chained gadgets could be executed. This would normally involve the use of a *ld* instruction to change the value of the previous return address to the next destination. For example, using *ld ra sp(40)* to load the malicious address from a stack under attackers control into the *ra* register to be fetched as next destination. If the gadget does not include this type of instruction, then the execution iterates over the last bunch of instructions via the previous value that is in the *ra* register causing a loop. In this case, execution results in a loop over the last bunch of instructions in the chained gadgets. The author in [17] also relate with a similar circumstance by giving another practical scenario where a bunch of code running on an electricity meter device should trigger a reload of credits to sustain service. However, an interference in transaction between user's bank and the meter, due to malicious code that leads to an endless looping of a bunch of code would not ensure that the meter is turned off if the user's account is not credited. This could be detrimental to the service provider as well as users.

There are various ways of evaluating new protections. The choice of percentage run-time performance evaluation is selected here because the executable binary changes after additional instructions have been added to it. A watchdog waiting time is included which inevitably increases run-time. In addition to these, the target now has to communicate with the monitor by passing out data via shared memory. All of these would impact the run-time as the program now does more than it was originally built to do. It is important to present the impact of the new technique with such useful detail so that the technique would adequately represent itself among other possible options. As producers of technological devices continue to build devices with variation in purposes, continuous study of possible protections for vulnerable software is needful. This continues to provide options in protections for users and vendors to choose from. This study has selected a distinct feature of watchdog waiting time combined with in-line CFI checks via inter-process monitoring as another means of enforcing CFI.

III. Threat Model

Previous studies have revealed ROP as a persisting threat to vulnerable programs. There could be other threats occurring in form of UAF [23] and double free [24], etc. This study focuses on ROP as a threat particularly when the chained ROP gadget ends up in a denial of service. The sample C program for this study was written with the bugs that are relevant for simulating the threat model. The program accepts input from user at some point in execution and also has a buffer overflow vulnerability which was leveraged upon to mount ROP attacks. Two new gadget finders were written to extract gadgets from the sample program and selected gadgets were chained to be passed into the target program as input to mount ROP attacks. The outcomes of the ROP attacks differed because of the difference in the ret gadgets and the order in which the gadgets were chained. A more detailed discussion on this would be featured in our future works. One of the outcomes from the various ROP attacks was selected for use here as threat model to demonstrate how the IP-CFI works. The gadgets were chained in a planned order such that when the byte stream is passed as input into the target process, a trap is hit where the program runs into a loop causing a denial of service.

IV. Summary of the IP-CFI Approach

In this section we discuss an overview of the IP-CFI approach. More details to elements in the protection system are given in Section V. The IPC-CFI is built as a protection system where the vulnerable process is monitored by another

process and values relevant for protection exist within the target process which is monitored by another process. The goal is to monitor the execution flow upon entry and exit from each function. The first step taken was to use GCC to compile the vulnerable C program with libraries inclusive so as to increase the possibility of having useful gadgets in the binary. Next, simple static and dynamic analysis on the vulnerable program to identify what elements are critical in the execution path of the process. Analysis tools were Ghidra for reverse engineering, Objdump for static analysis and Gdb for dynamic analysis. The analysis gave us a clear mapping of the intended control flow and the choice was made to insert lines of assembly instructions (enhancement code) to label all function prologues and epilogues.

The choice of *777*, *888* and *0* as values to be set as labels was made for experimental purpose and future works will introduce how the values could be hidden through encryption or applied at run-time. *777* is used to identify intended function prologues while 888 marks the unintended functions. The value *0* is passed at the epilogues to trigger a switch off in the in-line CFI value. The labels serve as values for in-line CFI checks as well as values for inter-process CFI checks. For inter-process, these values get written into shared memory through atomic operation of semaphore and mutex. The values are interpreted by the monitor as flags to indicate the status of the functions within the target during execution. This makes up the first part of the IP-CFI protection.

The second part of the protection involves the monitoring where the status value that was written to the shared memory is harnessed for further CFI checks. To achieve this, a C program was written which applies atomic operations to read into the shared memory. The program also implements a watchdog timer based on the status value read from the shared memory and halts the target process if a CFH is detected. In evaluating the effectiveness of the method, we analysed the run-time performance overhead. This was done by running two timed versions of the program with normal input 110 times. One version was the original program and the second version was one that had the enhancement code for protection applied before compilation. Data cleaning was done to eliminate 10 outliers form each data set and statistical analysis were done to validate the result of the two data sets of the run times in seconds. An average was calculated for each data set of 100 run-times. The average values were then applied to the formula:

Overhead = (Run-time with IP-CFI – execution time without IP-CFI)/execution time without IP-CFI

The overhead was calculated without the waiting time of 5 seconds which is required by the target process to enable inter-process communication. The overhead obtained is reasonable considering that some lines of code were inserted into the target for the new protection. The information obtained in the implementation were then used to make deductions and propose possible future works.

A. Exploitation and Protection Implementation Environment

The exploitation and protection implementation environment was set on a Linux Fedora computer system within which an embedded Linux Fedora RISC-V64 QEMU emulator (Fedora EM) was running as shown in Fig. 1. The use of the QEMU was necessary as the RISC-V architecture has not yet been adapted for direct run on PCs. The Fedora EM once started, was used to create, edit, compile and recompile the target program with all enhancements.

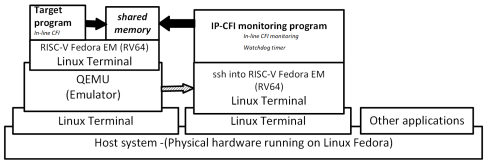


Fig. 1. Setting up a RISC-V System on QEMU Emulator

B. The Target Program

The target program used in this project is vulnerable to buffer/stack overflow and ROP. To simulate exploitation, we pass a of chain of ROP gadgets into the target. Details of this is given in Section V. Protection of the target in the first instance, adopts a CFI concept which makes use of checks inserted at the function prologue and epilogue of all direct functions. A comparison between the value that marks the intended function and the expected value from where it is called makes it possible for the CFI to detect a hijack. If the condition is not met, then the next instruction is executed, which halts the program, thereby avoiding further exploitation. As this In-line CFI on its own does not adequately protect the process from attacks stemming from ROP, we introduce other relevant protection concept in form of watchdog timing out where Inter-Process Communication (IPC) is used to establish IPM as a supervisory mechanism over the in-line CFI. In this case, the target process writes the in-line CFI value out into a shared memory and that value serves as the status value for an independent process to read and determine what action to take based on the status.

C. The IP-CFI Monitoring Program

The monitor is written in C program and values to be checked will be fetched from the relevant registers. It is an external independent process that is run concurrently with the IP-CFI-enabled target program to keep track of its execution. The external process consists of supervisory routine and a watchdog timing-out function which are implemented to ensure that the process maintains its intended flow. This Monitoring process communicates with the target program by reading its status from shared memory.

D. Shared Memory

In the Linux environment, there are two APIs that could be used to facilitate IPC- System V and POSIX. Both APIs provide IPC objects for reading and writing, but POSIX is safer to use as it does not permit execution for any category of user. According to [25], POSIX APIs are multithreaded-safe and we find it relevant for this project. POSIX APIs are also implemented with a backing file and we use that approach here to ensure compatibility, portability and persistence while the monitor accesses shared memory. In setting up shared memory, we mapped a shared file into the memory region *shm_open* using *mmap()*. The file could persist unless we delete it using

shm_unlink. We used *shm_open* to open the shared memory object and we used semaphore to avoid race conditions. The semaphore is also used as a mutex to lock/unlock access for the monitoring process and the target process.

E. Control Flow Integrity

CFI hinges on the ability of protection tool to observe the behavior pattern of the protected program during execution, detect anomalies and enforce the control flow integrity. The steps to it as implemented in IP-CFI is as follows:

1) Observing Program's Behaviour: The factors that are of importance in implementing this concept is a prior knowledge of expected flow of execution. This is a core step to the success of this method and it is achieved by carrying out thorough analysis on the vulnerable process. Each process varies in purpose, ability and vulnerability. The approach here is to establish the purpose of the process and identify vulnerabilities that are tied to the procedure by which that purpose is established. For example, a program that interacts often with users would have a higher attack surface. Also, processes that run for longer times will tend to be more vulnerable than short lived processes. The program is analysed by admin to identify the critical spots that lie within. The behavioural pattern is obtained from the CFG of the program prior to execution.

2) Detect Deviations: The success of a CFI-based protection depends on its ability to promptly detect a deviation from the expected pattern. Factors that are of importance here are the ability of the process to log in its status report into shared memory and to monitor delays in getting the status report value updated. The in-line value that is stored in the immutable registers are fetched and used to identify the status of the process. With this status, the watchdog is able to take the necessary action depending on the value read from shared memory. This also involves admin intervention prior to installation of the program. The time lapse that is permitted between the checks done by the watchdog is set prior to compilation.

3) Enforcing CFI: To enforce the integrity of the target process, CFI demonstrated here focuses on the external protection against CFH in a situation where attack bypasses in-line monitoring. This involves the insertion of instructions that enforce the CFI of the process. The enhancement code is inserted in assembly code into the protected binary. The needed elements for achieving these are relevant instructions and storage for the label values that are used in checking the legitimacy of each called function. The effectiveness of IP-CFI also lies in the trustworthiness of the detector/enforcer. To build an enforcer we combine two different sets of code in assembly language that are inserted into the target program - in-line CFI checks and Inter-process signaling code, and then a newly built external independent program that monitors the in-line CFI and inter-process status values.

F. The First Set of Enhancement Code

This fulfils the checks and enforcement of CFI and functions fully as in-line CFI. This relies primarily on the strategic positioning of checks in the target code. The positioning is determined by obtaining CFG of the program to see the possible pathway of execution. This involves a way of mapping out all subroutines in the program and identifying the direction of flow as intended by the programmer. A CFG could be built by making use of relevant reverse engineering tools. The authors in [4, 26] made use of Vulcan in developing a CFG but Vulcan is limited in use and is not implementable on RISC-V. Other useful reverse engineering tools are Ida and Ghidra which are effective in the x86 as well but none of these are yet to be effectively tailored to reverse engineer RISC-V-based applications. We have access to the sources code for our sample program here and that makes it a more straightforward process. However, in order to be able to use IP-CFI for programs that has been compiled without access to source code, a useful workaround that we applied was to use Ghidra on Fedora Linux running in x86 to reverse the x86 version of the same program. We found that this was useful foresight for easy analysing of the program in the RISC-V environment.

Further to this, static analysis informed us on the requirement for inserting the checks into the target. With these information, essential checks were inserted into the target in assembly code. The inserted in-line checks consist of fixed labels that mark intended functions along the execution pathway with matching values. Unintended functions along the execution pathways are marked with different values. At the beginning of execution, the value is set to *0* and would remain as 0 until a function call is made. If a function call is made, there would be two possibilities to the value which would either be a trigger to match the intended pathway or unintended pathway. The way labels are inserted would vary with the architecture of the underlying system. The RISC-V which is adopted here allows for straightforward storage of the values needed to achieve this process of the labels. The important objects for setting the values here are registers while the function prologue and function epilogues are used to position the in-line CFI checks and relevant action. Other pieces of code as shown in Fig. 2 that work in this phase are geared towards halting the target process based on the in-line CFI checks.

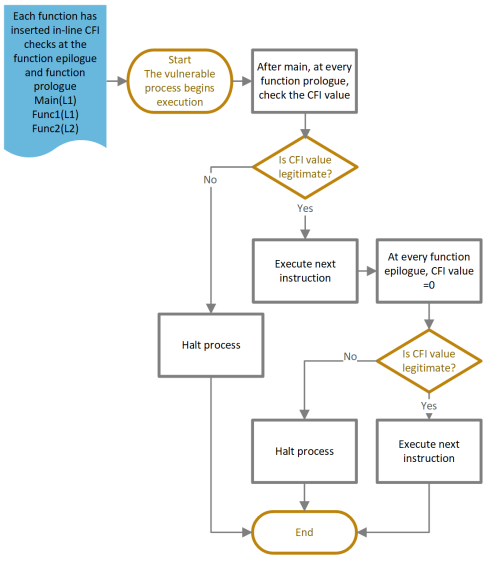


Fig. 2. Flow Chart of Target Process with In-Line CFI.

1) Wrongmove: This is a two-line instruction block that performs the halt to the vulnerable process by making a system call. For a sensitive program like the target, a proper handler needs to be triggered once a CFH is detected. Hence for this protection, the in-line CFI hands over control to the operating system by making an *ecall* on RISC-V. The *ecall* is an established RISC-V system call that ensures the transfer of control to the kernel so that the attacked process can be halted. This is achieved by setting the relevant value of *93* into a register *a7* in assembly code.

2) Storing the Label: The values used in the in-line CFI are set as numbers *777* and *888* to demonstrate the protection as these could be other values, provided that the value held by each of the intended nodes match the legitimate caller. An important decision that was made here is where to store the value to ensure that it is preserved throughout each function.

In RISC-V, values are stored in registers but the important register for storing our values were selected based on specific attributes. As the value in the label will be used for monitoring control flow status, it is important for the value to be unchangeable within its function. For this reason, the *s2-s11* registers are found suitable. These registers are referred to as saved registers. They have the ability to preserve the values stored in them within the function. Subroutines do not normally change the values and If they do, they will have to save the value and restore at the end of its execution.

For this purpose, only registers in this category can be used as other registers do not share the same attribute. They will either have specific roles or are temporary registers which means that the value that they hold will not be preserved. Apart from preservation, it is important and ideal to have labels that cannot be manipulated by attackers. This limitation is being studied for future improvement on the protection method. This would make those registers a strong tool for the implementation of this protection technique on RISC-V. For the demonstration here, the *s3* and *s4* registers were selected. Another option to using a known value as the label is to generate a scrambled value at run-time. This is however outside the scope of this study but is an area that could be considered for future studies. Outline of in-line checks as as follows:

G. The Second Set of Enhancement Code

This fulfils inter-process communication by logging status report from the target program to a shared memory. The status report is fetched from the outcome of the in-line CFI mechanism and values held in labels that have been set to mark the main execution path of the process, taking cognizance of its entry into and exit from critical nodes. This code is a new function that consistently writes the value contained in the *s3* and *s4* registers into a log to share the status of the target with a monitoring process.

1) Status Logging: The instructions that handle this step is inserted into the target assembly and it carries out *open*, *write* and *close* system calls to achieve this. It also applies an atomic operation involving a semaphore to these calls to avoid race conditions. Portions of this code were retrieved from [27] examples of shared memory.

H. IP-CFI

The CFI monitoring is enhanced by attaching an additional monitoring method involving another process attributed as not vulnerable as it runs with zero user interaction. The monitoring process performs the function of observing the target process by implementing an atomic inter-process signaling. The main tool that the monitor uses is information read from memory shared with the target process. In the case of ROP attack, it was observed that the status could appear legitimate whereas, a hijack has occurred undetected. For this reason, monitoring is extended to watchdog timing out.

1) The Watchdog Routine: The watchdog routine sets a counter to keep track of the target process. The demonstration in this report gives allowance of three cycles of checks by the watchdog. The first and second cycles could be enabled to restart the process without user intervention while the third cycle puts a halt to the target process. The number of allowable cycles can be adjusted to suit the performance or function of the protected program. For the demonstration here, the restart is not included for any of the cycles.

V. Implementation

As this protection is aimed at combating memory compromise through buffer/stack overflow that lead CFH via ROP input, promptness and accuracy in detecting deviations is very important it. The earlier a protection system is able to detect deviation and enforce integrity, the higher the chances of establishing a secure process.

A. Exploiting the Target

The first step to demonstrating the protection is to demonstrate an exploit. ROP is dependent on availability of gadgets and the ability of attacker to craft a byte stream to accomplish their malicious goal. Implementing ROP on RISC-V is more complex than x86 but is achievable. The aim is to control the execution from the stack by passing carefully crafted input through buffer overflow.

B. Gadget Finders

In order to make a variety of gadgets available to us, we wrote two gadget finders, *RETGadget* and *JALRGadgets* in Linux scripts and applied with the target as an argument to extract gadgets from it. The gadget finders are available and can work on RISC-V-based programs.

C. Passing Chained Gadgets

Once the gadgets were extracted, we mapped out ROP chains in various order based on a theoretical approach by [28]. According to [28], we can pass gadgets that will help us to store values and addresses in the registers that we intend to use to mount the ROP. The author in [28] classified the gadgets as functional gadgets and charger gadgets. The functional gadgets will hold the instructions for the actual attack while the chargers(linkers) gadgets will load the registers with addresses of the functional gadgets and other useful values. The linkers can be used to create a fake frame as shown in Fig. 3 and 4, which we can exploit further to pass malicious values into registers and other elements on the stack. Each of the ROP

chains was passed as input into the target to see how it could be exploited. In Fig. 3, the charger simply creates the fake frame with values loaded or copied from one location or register to the other. An exit is then called.

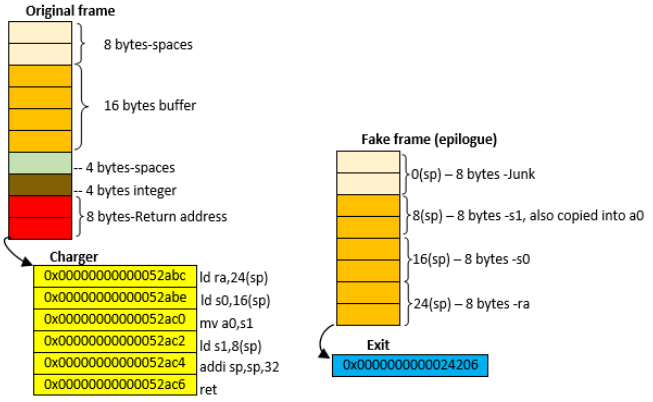


Fig. 3. Exploiting Target to Manipulate Some Registers and Exit Abruptly.

Further exploit is done as shown in Fig. 4 where two functional gadgets are chained to the linker. However, the outcome of this chained gadgets is different from that in Fig. 3 as this results in a loop. This is because we used a functional gadget that does not overwrite its previous value in *ra* register.

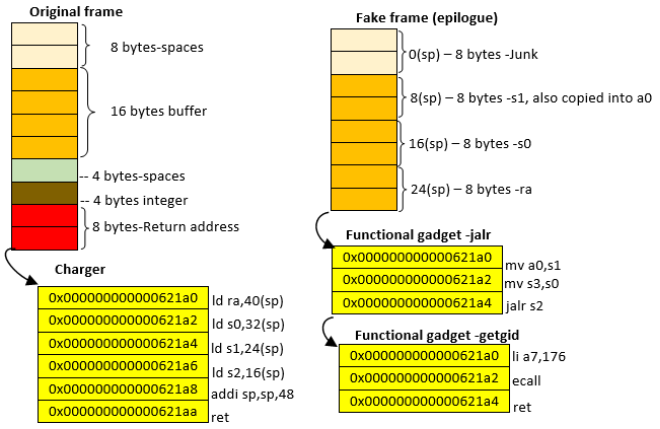


Fig. 4. Exploiting the Target to Cause a Loop.

The order in which gadgets are crafted in RISC-V would determine the outcome of the exploit. This has a lot to do with the value of the *ra* register that gets overwritten from time to time as execution steps into and out of library functions or other functions that get called within a function. This can only be detected during dynamic analysis and remains undetectable to user as no feedback is written to standard output but the process appears to be hanging as it doesn't crash. This a typical attack that could lead to denial of service and the IP-CFI protection is able to detect and handle it.

D. Protecting the Target

With the attack in place, we then applied the IP-CFI protection. The monitoring process is run concurrently with the target. Each time a new function is called in the target, the status is updated by the target process via the shared memory as shown in Fig. 5. The status value indicates what sort of function is being executed and at what stage of the function the execution is. Once the monitor reads into the shared memory, it

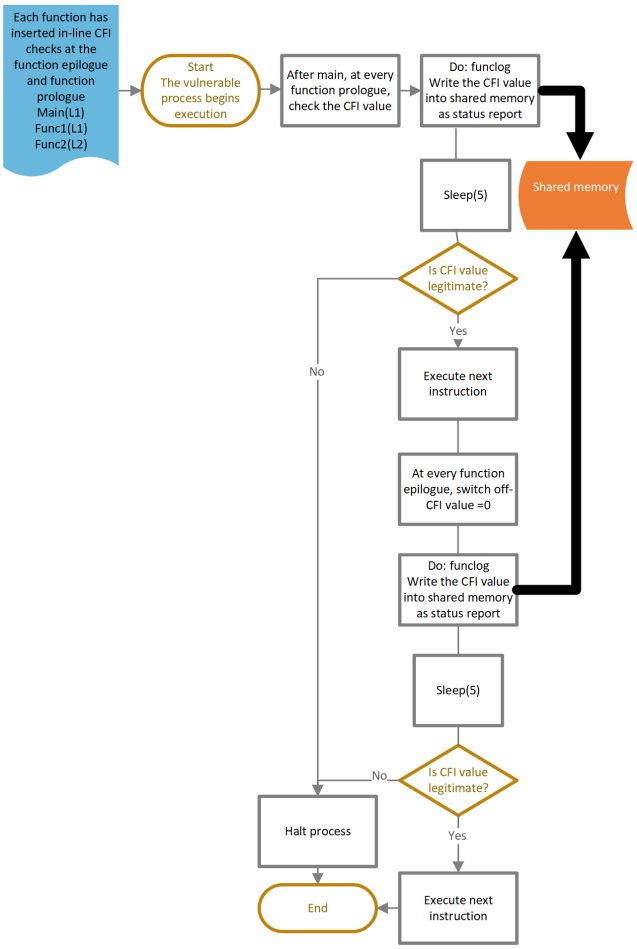


Fig. 5. Target Writes into Shared Memory.

takes the necessary action depending on the value. The denial of service is promptly identified and stopped. The possible flow of execution of the monitor based on the possible values that could be held in the status report is shown in Fig. 6.

E. Applying IP-CFI to a Source Code

There are three steps in setting up a program to use IP-CFI. We begin with the C source code and end up with an executable. The steps include two stages of compilation with the insertion of enhancement code between stages. The relevant scripts:*IP-CFI-make.sh* and *IP-CFI-full-compile.sh*, the monitor program *IP-CFI-monitor-watchdogv1*, and the enhancement code *IP-CFI-enhancement-code.s* are required.

Step 1:
Run *IP-CFI-make.sh* passing the C source code as argument

Step 2:
Find the resultant assembly (.s) and manually insert CFI-checks. Instructions are mapped out as follows:

Within main function:
Insert the following lines before the first call to a function

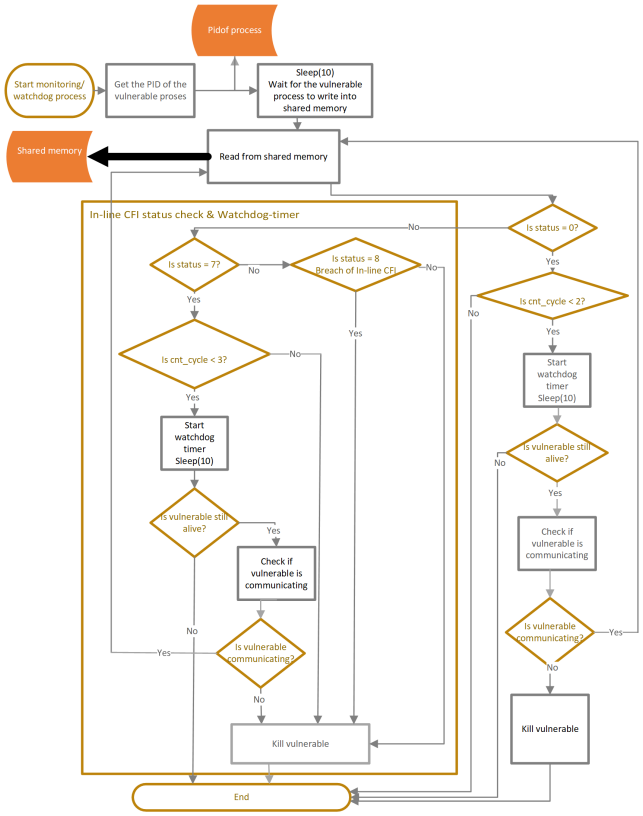


Fig. 6. Flow Chart Showing Inter-Process Monitoring as Status is Read from Shared Memory.

li s3,777 #set monitoring label for legit function

Within all intended functions that are not main:
Insert the following lines at the prologue:

Line 1:

li s4,777 #inserted for legit function

Line 4:

call IPCFIfunclog #log status

bne s4,s3,.Wrongmove. #enforce in-line CFI

Insert the following lines at the epilogue (just before *ra* register is loaded or before call to *exit*):

li s4,0 #Switch off CFI label

call IPCFIfunclog #update status

Within all unintended functions:
Insert the following lines at the prologue:

Line 1:

li s4,888 #inserted legit label

Line 4:

call IPCFIfunclog #log status

bne s4,s3,.Wrongmove. #enforce in-line CFI

Once all checks have been inserted save the file and exit.

Step 3:
Run *IP-CFI-full-compile.sh* passing the saved assembly (.s) file as argument. The resultant executable within current working directory can be run concurrently with the monitor. Work is in progress to fully automate these stages.

VI. Evaluation

In evaluating the protection, further observations were made with various forms of ROP chains passed into the target. An exploitable target was used. With the watchdog in place, the protection was applied to the target while different ROP chains were passed as input into it. The protection was found to be effective. The impact that the additional prologue and epilogue code might have on the execution time of the program was considered. In this section, we evaluate the new protection by obtaining the relative performance overhead with respect to our target program.

A. Results and Discussion

The outcome of analysing the run-time overhead of the new protection is discussed in this section. While IP-CFI effectively surmounts ROP with a run-time overhead of 1.5%, there could be variations in the outcome execution time based on the waiting time set in order to accomplish synchronisation between the monitor and the protected process. Details of how the run-time overhead is calculated are as follows:

Run-time type	***Average of run-time***
w/o IP-CFI	*3.26 ± 0.12 (milli seconds)*
w/ IP-CFI	*8.06 ± 0.16 (milli seconds)*

Run-time overhead w/ IP-CFI

((8.06 – 3.26)/3.26) = 1.47239264 (approx. 1.5%)

This presents a reasonable overhead when the additional pieces of code are included into the target. When the program is run concurrently with the monitor, some waiting time is required in order to establish communication between the target and monitor. This could vary from one program to another as it largely depends on the purpose of the program. For the sample program, a 5 seconds waiting time is applied to accommodate the time required for the monitor to read into the shared memory and take the necessary action.

While the full protection surmounts ROP, the waiting time applied appears to significantly increase the overall response time. However, the extra time incurred here is artificial. In reality, it is not additional run-time as the program would function fully without the waiting time. Waiting is needful to achieve interoperability between the two programs here and this outcome could differ if various scenarios are considered. In this instance, IP-CFI has been applied on a simple program and the outcome may vary more favourably with larger programs. On the path of the monitor, a 10 seconds wait is involved but this is independent of the target and does not impact the target run-time.

Furthermore, the run-time may vary slightly with the number of functions that accept the enhancement code. However, since the functions will only run one at a time, the overhead would not be greatly impacted. Also, optimization was set to *0*for the samples used here. The two options of level of protection could be applied to vulnerable process- one with full IP-CFI, and the other without the watchdog waiting time.

An alternative to the IPC via signalling, in order to monitor in-line CFI with less waiting time is the direct monitoring of the process from the kernel. A study by [31] presented this as a security measure but not with regards to CFI. According to [31], every process reports its logs somewhere in the file system and this could be harnessed as useful information for the kernel to monitor processes. If the kernel were to be used directly for monitoring the CFI, other related elements like the in-line CFI value and watchdog timing routine might need to be reconsidered.

A limitation of IP-CFI is that the values in the labels might leak and the registers that holds those values could be reused by an attacker. Although the impact of this with IP-CFI implementation against a CFH does not yet appear to be considerable. The possibility of locking the s3 and s4 registers is being considered for future works, as well as ways to encrypt the label values to generally improve on the IP-CFI protection.

Currently, RISC-V applications exist in highly sensitive eco-systems as they are commonly used and constantly running. IP-CFI is aimed as a broad spectrum of protection cutting across various eco systems. For efficiency, it is however aimed to protect applications that are built for long running services or those performing a single role. The RISC-V platform is well structured for such applications and is expected that the intended purpose of the application would inform the choice of protection. The RISC-V architecture is being implemented for several health monitoring devices. A recent study by [29, 30] presents a cutting-edge technology in form of an implantable medical device (IMD) for conditioning the human body electrical activity which runs on a RISC-V processor. [11] also produced a RISC-V-based microprocessor that could be used in devices for personalized health management aside from other devices like electronic voting machines, smart cards, etc. These devices match the category of devices that are dedicated for a single purpose and applications that are run on them might benefit from the IP-CFI protection.

VII. Conclusion and Future Works

Here, we have presented a proof of concept using IP-CFI, a new protection mechanism which is based on the concept of CFI combined with an external monitoring program. IP-CFI effectively resolves a denial of service from lingering when ROP is mounted. The main strength of the system is its ability to detect delays in change of the status value logged into shared memory where the monitoring process fetches information for taking actions towards maintaining the integrity of the protected program. With a prompt detection of delay, the target process can be halted to prevent furtherance of attack process.

The possibility of sustaining an execution while preventing furtherance of attack is an area that previous CFI solutions have not really addressed as CFI tends to halt processes once an attack is detected. One of the areas we explored in the process of this study is a way that the monitor could trigger a restart to the program rather than a halt. So far, we have no way of preserving existing data such that the restart of the process is done without side effects. This option would be explored in future works especially for environments that some of these vulnerable processes might require seamless continuity.

In this study, we have opted for a higher-level monitoring process to give us more control of the protection, as well as increase flexibility in the settings. The overall response time can be improved upon by optimising the IPC and setting the monitor to respond asynchronously. This is being considered for future works.

References

[1] H. Shacham, "The geometry of innocent flesh on the bone: return-into-libc without function calls (on the x86)," ACM conference on Computer and communications security, pp. 552-561, 2007.

[2] Microsoft Corporation, Microsoft Documentation, 31 May 2018. [Online] Available:https://docs.microsoft.com/en-us/windows/win32/memory/data-execution-prevention.[Accessed 22 April 2020].

[3] Tran, M., Etheridge, M., Bletsch, T., Jiang, X., Freeh, V., Ning, P, "On the Expressiveness of Return-into-libc Attacks," Sommer R., Balzarotti D., Maier G. (eds) Recent Advances in Intrusion Detection. RAID 2011. Lecture Notes in Computer Science,, vol. 6961, 2011.

[4] M. Abadi, M. Budiu, U, Erlingsson, and J. Ligatti, "Control Flow Integrity," in CCS '05: Proceedings of the 12th ACM conference on Computer and communications security, New York, United States, 2005.

[5] B. &. T. G. Niu, "Modular Control-Flow Integrity," in PLDI '14: Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation, 2014.

[6] De, A, Basu, A, Ghosh, S.,& Jaeger, T., "FIXER: Flow Integrity Extensions for Embedded RISC-V," in 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2019.

[7] V. Kuznetsov, L. Szekeres, M. Payer, G. Candea, R. Sekar & D. Song, "Code-Pointer Integrity," in The Continuing Arms Race: Code-Reuse Attacks and Defenses, Association for Computing Machinery and Morgan & Claypool, 2018, p. 81–116.

[8] C. Sadullah, D. Leila, Z. Boyou, J. Ajay & E. Manuel, "Efficient Context-Sensitive CFI Enforcement Through a Hardware Monitor," in Detection of Intrusions and M17th International Conference, DIMVA 2020, Lisbon, Portugal, June 24–26, 2020, Proceedings, Lisbon, Portugal, 2020.

[9] M. Neugschwandtner, C. Mulliner, W. Robertson, & E. Kirda, "Runtime Integrity Checking for Exploit Mitigation on Lightweight Embedded Devices," International Conference on Trust and Trustworthy Computing, vol. 9824, pp. 60-81, August 2016.

[10] A. Waterman, K. Asanovi & J. Hauser, "The RISC-V Instruction Set Manual, Volume II: Privileged Architecture, Docu- ment Version 2021120," 2021.

[11] IIT Madras, "IIT Madras, Indian Institute of Technology Madras," IIT Madras, 24 Sept. 2020. [Online]. Available: https://www.iitm.ac.in/happenings/press-releases-and-coverages/iit-madras-develops-and-boots-moushik-microprocessor-iot. [Accessed 08 July 2022].

[12] DeepComputing, "xcalibyte.com," xcalibyte, 28 06 2022. [Online]. Available: https://xcalibyte.com/roma-preorder/. [Accessed 04 07 2022].

[13] A. Samuel O, "An Overview of RISC Architecture," in Proceedings of the 1992 ACM/SIGAPP Symposium on Applied Computing: Technological Challenges of the 1990's, Kansas City, Missouri, USA, 1992.

[14] G, Gu, and H. Shacham, "No RISC No Reward:Return-Oriented Programming on RISC-V," 29 July 2020.

[15] G-A. Jaloyan, K. Markantonakis, R. N. Akram, D. Robin, K. Mayes, and D. Naccache, "Return-Oriented Programming on RISC-V," ASIA CCS '20: Proceedings of the 15th ACM Asia Conference on Computer and Communications Security, p. 471–480, October 2020.

[16] T. Oyinloye, L. Speakman, T. Eze, "Inter-Process CFI for Peer/Reciprocal Monitoring in RISC-V-Based Binaries," in 20th European Conference on Cyber Warfare and Security, 2021.

[17] F. Stajano, R. Anderson,AT &T Laboratories Cambridge, "The Grenade Timer: Fortifying the Watchdog Timer Against Malicious Mobile Code, " in Proceedings of 7th International Workshop on Mobile Multimedia Communications (MoMuC 2000), Waseda, Tokyo, Japan, 2000.

[18] S. Sayeed, H. Marco-Gisbert, I. Ripoll, and M. Birch, "Control-Flow Integrity: Attacks and Protections," Applied Sciences, vol. 9, no. 20, p. 4229, October 2019.

[19] M. Zhang and R. Sekar, "Control flow integrity for cots binaries," SEC'13: Proceedings of the 22nd USENIX conference on Security, p. 337–352, August 2013.

[20] K. Onarligolu, L. Bilge, A. Lanzi, D. Balzarotti, and E. Kirda, "G-Free: defeating return-oriented programming through gadget-less binaries.," in Proceedings of the 2010 Annual Computer Security Applications Conference, New York, NY, 2010.

[21] A. Srivastava, A. Edwards, and H. Vo., "Vulcan: Binary transformation in a distributed environment," Microsoft Research:Technical Report: MSR-TR-2001-50, 2001.

[22] De, A, Basu, A, Ghosh, S., & Jaeger, T., "Hardware Assisted Buffer Protection Mechanisms for Embedded RISC-V," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2020.

[23] Common Weakness Enumeration, "cwe.mitre," 19 July 2006. [Online]. Available: https://cwe.mitre.org/data/definitions/416.html. [Accessed 17 July 2022].

[24] Common Weakness Enumeration, "cwe.mitre," 19 July 2006. [Online]. Available: https://cwe.mitre.org/data/definitions/415.html. [Accessed 17 July 2022].

[25] Oracle, "Oracle Programming Interfaces Guide," 2012. [Online]. Available: https://docs.oracle.com/cd/E26502_01/html/E35299/svipc-posixipc.html#scrolltoc. [Accessed 04 06 2022].

[26] E. Göktaş, E. Athanasopoulos, H. Bos, and G. Portokalidis, "Out Of Control:Overcoming Control-Flow Integrity," in 2014 IEEE Symposium on Security & Privacy, San Jose, CA., USA., 2014.

[27] M. Kalin., "Inter-process communication in Linux: Shared storage," 2019.

[28] B. Deac, "InfoSec Write-ups," 14 March 2022. [Online]. Available: https://infosecwriteups.com/return-oriented-programming-on-risc-v-part-1-dd9817b52d2b. [Accessed 25 06 2022].

[29] A. Arnaud, M., Miguez, J. Gak, R. Puyol, R. Garcia-Ramirez, E., Solera-Bolanos, R. CastroGonzalez, R. Molina-Roblkes, A. Chacon-Rodriguez, R. Rimolo-Donadio, "A RISC-V Based Medical Implantable SoC for High Voltage and Current Tissue Stimulus," in 2020 IEEE 11th Latin American Symposium on Circuits & Systems (LASCAS), Costa Rica, 2020.

[30] A. Arnaud, M., Miguez, J. Gak, R. Puyol, R. Garcia-Ramirez, E., Solera-Bolanos, R. CastroGonzalez, R. Molina-Roblkes, A. Chacon-Rodriguez, R. Rimolo-Donadio, "Siwa: a RISC-V RV32I based Micro-Controller for Implantable Medical Applications," in 2020 IEEE 11th Latin American Symposium on Circuits & Systems (LASCAS),, Costa Rica, 2020.

[31] W. Kehi, G. Yueguang, C. Wei & Z. Tong, "The Research and Implementation of the Linux Process Real-Time Monitoring Technology," in 012 Fourth International Conference on Computational and Information Sciences, 2012.

Towards Flexible Transparent Authentication System for Mobile Application Security

Abdullah Golam, Mohammed Abuhmoud, Umar Albalawi
College of Computing and Information Technology, University of Tabuk, Tabuk, 71491, Saudi Arabia

***Abstract*—Undoubtedly, Mobile Application Security (MAS) has made tremendous progress in implementing enhanced security protocols in the past decade. With the recent increase in the usage of mobile applications, concerns of privacy and security are increasing rapidly. Thus, the security measurement must be applied to satisfy security and privacy needs. On the one hand, the developer community works feverishly to develop mobile applications with innovative and usable layers and user-friendly for multigenerational customers. However, the security community, in particular, strives to make those layers secure. Therefore, the main objective of this research is to build a transparent authentication system in a mobile application. There are potentially many ways to implement an authentication mechanism such as the biometrics approach. It has features, which can be used to heightened security for the end-user. In these articles, we experimentally investigate the multigenerational customer base's factors such as age, convenience, easiness, memorizing new passwords, and understanding the precept of frequently changing passwords to enhance security. Additionally, we propose a system that will solve the common problems users face when starting the password resetting process. At the same time, in the MAS sector, we orchestrate the applications for better security encryption for the stored biometrics to ensure it, which makes it even more challenging for an adversary to bypass the system and reset the password. We conclude our research with a comprehensive security solution for MAS that considers user friendliness and data safeguarding.**

***Keywords*—Transparent security; authentication; UX/UI; forgetting password; reset password; biometric systems**

I. Introduction

Considering the tremendous development in the use of mobile applications and the search for hacker-proof programs that are gathering momentum, security problems have become a feature in the developer's mind to the point of obsession. This endeavour goes hand in hand with the need to enhance the user experience through user-friendly products in times of unprecedented demand for programs that combine security with simplicity and fun. To this effect, the duality 'Security' and 'Usability' operate as an interdependent set of elements, which the developer must incorporate with equal measure. The implications of failing this key operational balance between "Security" and "Usability" will most likely lead to a lack of security or to some difficulties in using the application.

For almost two decades, authentication has been a prominent issue in usable security research. The majority of these studies have concentrated on passwords and other similar authentication techniques that rely exclusively on a shared secret between the user and the computer system. Passwords must be strong enough to prevent guessing and must be memorable enough to not need to be written down. They must also be easy to type and lengthy enough to prevent guessing [1].

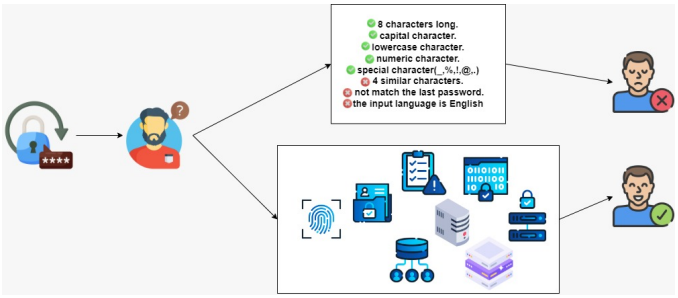


Fig. 1. Password vs. Fingerprint in an Authentication Process.

Transparent security related to mobile authentication can be implemented by physiological biometrics especially using fingerprint. In fact, the fingerprint is rare and difficult to obtain without the consent of the owner. It is also easy to adopt the fingerprint in giving access to the user with authority, as it is considered one of the most popular modern methods, which have been conducted multiple researches measuring the feasibility of this method in terms of ease of use and certainly security [2].

In Fig. 1, it is clear how the traditional method is complicated in resetting the password. For example, the user must not use the same password as the previous and that it is not identical to the name of the user and not to his personal data like date of birth. Also, it must contain symbols, uppercase and lowercase letters, and be at least eight digits long. On the other hand, you only need to pass your finger over the fingerprint sensor and in light of the unique data of the fingerprint. It is safe and has reliable features in its use to give access, Thus, using fingerprint in authentication satisfies the security measurements in a transparent manner. In most cases user do not feel what is happening in the background The system consists of very complex calculation in image processing and fingerprint Features extraction to verify the fingerprint provided by the user to give him access [3].

Although it is not in a user's interest to forget his password, many have done so. The age factor and how it affects the user's memory box, or the need to create as complicated a password as possible to keep the adversary at bay can inadvertently cause a person to forget their password. Either because we may have forgotten our password or due to a security breach, we find ourselves compelled to. As cybersecurity researchers, we must

use all possible means to make this process as secure and user-friendly as possible. Using biometrics that cannot be forgotten or stolen like a security token is one way, the other entails adding another layer of security encryption which must be applied on the stored biometric to ensure it will be even harder to bypass the system and reset the password. As a solution to this ongoing issue, we propose a system that will solve the common problems users face when starting the reset password process.

II. Background and Related Work

A. Transparent Security

Transparent authentication systems for mobile devices can be classified based on whether they use physiological biometrics such as fingerprint scanning or face recognition, or behavioral biometrics such as keystroke touch or walking rhythm. Physiological biometrics are widely regarded useful because they require a lot of computational power and high-quality photos, which are difficult to breach. Iris recognition, for example, requires the user to face the camera, takes longer to authenticate, and requires expensive additional hardware. Furthermore, iris recognition still faces obstacles such as detection, segmentation, coding, and matching [2]. Fingerprint recognition, on the other hand, with the progress of the smart phone industry and the adoption of technologies that provide users with a unique experience for their devices, does not face these obstacles. We find that most modern devices include a fingerprint sensor, and from this comes our focus in this paper on using this sensor to build a password reset system using the fingerprint, which is considered a transparent authentication system.

B. Usability vs. Security

Usually, the security part is sacrificed to complete the user experience part, and vice versa. The proposed idea is a balance between the two, leading us to the term "Transparent Security" intended to complete the security process so that the user does not feel the existence of complex security operations and where some waiting is required for the completion of these processes, usability techniques can be used to make waiting not boring by using usability emotional design. If the system is difficult to use, users will avoid using it. It must be taken into consideration that the effective use of applications and programs requires the programmer to implement usability while designing any program because that will affect efficiency and performance when using the program. The design of effective applications should consider the language, cognition and social interpretations of the user and the community. The word "Usability" also refers to methods for improving ease-of-use during the design process. Usability is defined by five quality components: learnability, efficiency, memorability, error, and satisfaction [4].

Security and usability are acknowledged as working in conjunction. There are examples of security and usability disputes, and these involve: password creation complexity instructions which will be hard to memorize, the enforcement of password masking to save passwords from being compromised, which sacrifices usability [5]. Information security is the defence of individuals, communities, or national interests, along with their information and noninformation-based properties, from the risks associated with their interactions with cyberspace. Users and their communities are among the properties that must be protected. Several security professionals and countries are now recognizing the need for users to be more informed and informed about information security [6].

C. Forgetting Problem

Forgetting is part of contextualization and guides immediate and potential information processing by encouraging environmental exposure and ensuring that knowledge is up to date, enabling timeliness and up-to-date [7]. The problem of forgetfulness cannot be overcome, as it is part of human nature. Therefore, different application developers must consider this human characteristic. As a result of Carnegie Mellon University password security research [8], strong passwords are not easy for users to implement and memorize, the problem is aggravated by users needing to implement and memorize special passwords for all online accounts they use. Joseph Bonneau and his colleagues evaluated 20 years trying to find a password-alternative proposition. They created a collection of 25 criteria that concerned usability, security, and deplorability and used them to assess different authentication methods. They concluded that there are no password alternatives that offer many advantages over conventional passwords. Furthermore, many did not meet a sufficient range of real-world constraints as password alternatives.

D. Password Resetting

Among the things used to reset the password is CAPTCHA [9]. This requirement is meant to ensure that the user is a real person, but the system is greatly affected by the user's experience in terms of clarity and ease of reading and may face challenges in determining the content displayed in front of him/her. These challenges sometimes impede the user from passing the selection point or force them to spend considerable time trying to pass beyond that identification test. Let us not forget, users may be of different ages, may have a low level of education and do not understand English, the preferred language by developers in use by CAPTCHA. The presence of vague characters is also another type of word blindness, which also leads to the character not being specified due to the overlap between the letters, and many of the examples are mentioned in discussing this method. However, the problem in terms of usability refers to the oversight of developers in failing to address the issue of age when designing their systems. The implications indicate neglect of certain age groups. For example, if the user is in an advanced age group, can CAPTCHA be made less challenging than the one designed for younger users? Moreover, there lies a challenge and what if the user is of Arab stock with no knowledge of the English script and is yet expected to answer questions written in English. Here lies a challenge that is too great to overcome. The main problems in this method can be summarized as follows:

1) Distortion issues.
2) Content issues.
3) Presentation issues.
4) Location and position.

The graphic design [10] of the password is used to make it more memorable, user-friendly, and secure. Put simply,

TABLE I. COMPARISON BETWEEN FINGERPRINT AND FACE RECOGNITION

Biometric System	Fingerprint	Face Recognition
Universality	Medium	High
Uniqueness	High	Low
Permanence	Medium	Medium
Collectability	High	High
Performance	Medium	Low
Acceptability	Medium	High
Circumvention	High	Low

the user presses the shape based on which he created the password and then logs into the system. The focus was on usability, and the goal was to test the user experience for the picture password. The questionnaires were used as a tool that covered many users of different categories, including age groups and cognitive achievement. The main argument for graphical passwords is that humans are better at memorizing graphical passwords than alphanumeric character passwords.

III. WHY FINGERPRINTS

There are seven features that determine biometric advantages: universality, uniqueness, permanence, collectability, performance, acceptability, and circumvention [11]. The concept of universality states that we can always be successful in finding our desired biometric features in the number of people who will be enrolled in the system. The term uniqueness refers to the number of distinct features a person has among people. Permanence is the evaluation of how far a unique set of characteristics endures or variates over time with maturity level. The concept of collectability refers to measuring how quick and easy it would be to obtain features that can be used to verify identity. Performance refers to a collection of measurements used to evaluate how well a given set performs. Speed, accuracy, and error rate are examples of these measurements. The acceptability of using biometrics evaluates how adequate and satisfiable it is. Circumvention refers to the ease with which a system can be fooled by a forged biometric feature. In mobile devices, the most widely used biometric sensors are fingerprint and facial recognition sensors [12].

The fingerprint is unique even in twins, it will be different and is distinct from one person to another. However, it can be forged like a dummy finger, new technologies have emerged which eliminate this problem by adding some pulse detector and temperature sensor. The fingerprint is a suitable biometric system, as shown in Table I, because it is hard to collect without user cooperation. In face recognition, it has a uniqueness problem due to the similarity between siblings, especially in a twins situation, and that will cause a problem if someone tries to access a system using face recognition like twins. Bad performance depends on many factors, such as the accuracy and speed to analyze. One of the problems of face recognition is that any person may get the face template of someone else from a far distance by using a super-zoom camera that captures long-distance shots, which causes alarming concerns. On the basis of these facts, it is suitable to choose a fingerprint as a biometric system in our proposed system because it will add a security layer and be easy to use and store.

IV. USER EVALUATION OF CLASSIC PASSWORD RESET

Interviews were conducted to evaluate the classic resetting password for various systems as presented in Fig. 2. Each system was developed with different types of password reset methods. The target participants we had interviewed use internet services for various purposes. Most of the usage orbits around browsing to benefit from the services provided on the Internet and to communicate with family and friends through social networking applications. Since social networking sites and other websites store cookies on user devices, the user does not log in by entering his/her username and password, often which means forgetting what they are. Participants in the interviews indicated that they may periodically reset the password, every six or twelve months, and look at the expiration date of the cookies. Therefore, most sites usually leave the configuration for this feature as default, which indicates the period for storing the cookies from six months to a year. Another factor that may cause a password to be forgotten is a different password for different platforms. To avoid the forgotten password, as most of the participants stated that they use the same password on more than one platform and app.

We want to explore the experiences of technical and non-technical Internet users with existing internet services for various uses. The goal is to determine what conditions may require the user to reset their password and what will make this process easier and more secure while saving the user time. The findings of the interviews conducted are as follows:

- The factors that motivate participants to change their password: some participants who may be either technically oriented users or plainly nontechnical change their passwords because they merely memorize their password to keep it safe.
- Degree of satisfaction with the reset password process (0-10): Most participants rated the process of resetting a password below 8 and this tells us that they did not reach their expected level of satisfaction due to the complex methods of resetting password, they feel the operators subject them to.
- The length of time it takes participants to reset a password: The method used in such a process consumes longer time, than the generally expected norm. This depends on the mechanism of resetting passwords, which affects the efficiency even though users understand the importance of the long-time process.

V. THE PROPOSED FLEXIBLE TRANSPARENT AUTHENTICATION SYSTEM

In the proposed system, our goal is to improve the quality of the user experience by using modern techniques such as biometric systems. In the sign-up phase, the system will ask a new user to enter their name, email address, mobile number, password and fingerprints, as shown in Fig. 3 and the Algorithm 1.

In the in-session log-in, it will ask for two things, the email address of the account and the password to gain access to it. If the user has forgotten his/her password, the user will select the forgot password batten. In this phase, there are few fields to complete. First, enter the account email address, then the

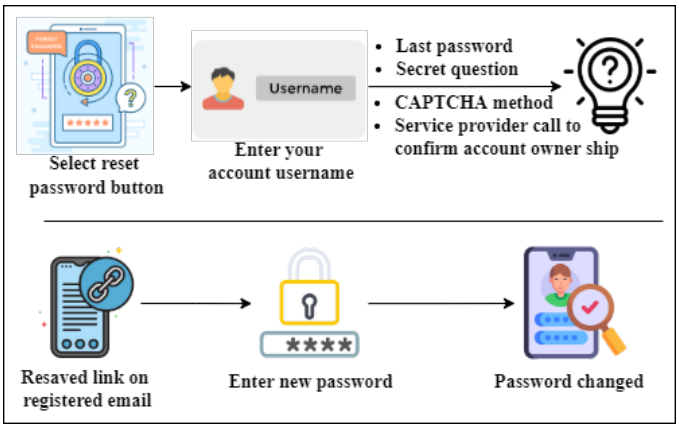


Fig. 2. Classic Password Reset Process.

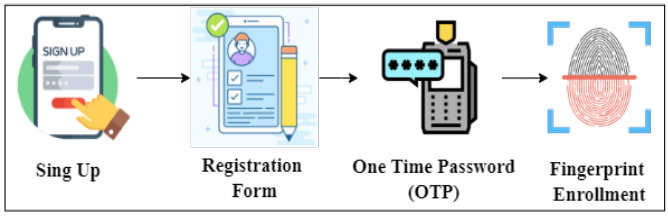


Fig. 3. Sign-Up Process.

user will get an OPT code (a temporary code sent to the user's phone) to confirm that the person trying to reset the password is the owner of that account. After that, the user will have the choice to choose from two methods to reset the password:

1) Fingerprint: if the user chooses to reset the password using the fingerprint method, the user will place it on the scanner, for the system will match the template. If the template matches, then the system will directly open the reset password page, explained in Fig. 4 and Algorithm 2.
2) Email address: if the user chooses the email address, the system will send a link to the user's email to reset the password. For a degree of flexibility that will support both usability and security, the proposed system has two methods for the user to choose from. It eliminates considerable problems by adding security layers and improving usability. Users may suffer from blurred vision, as in entering vague characters during the CAPTCHA test to reset password. Even users who have no blurred vision problem will feel discomfort when trying to enter vague characters during the CAPTCHA phase.

Biometrics, especially fingerprints, are widely used among other biometric systems in authentication tests. Most systems save the user's unique template of minutiae directly in the database as a special template for the user, which can be exposed to a possible attack. This unique and limited information is exposed to danger, as it is possible to reconstruct the fingerprint from the leaked information. To this end, there is a need to urgently enable protection of this information [13].

The hash algorithm is a complex mathematical function that transforms a collection of inputs into an apparently random output string of fixed length, so the same input string will always produce the same output string [14]. However, if the input string has changed even by just a single character, then the output string will be entirely different. Ordinarily, encryption implies incidentally scrambling information until a key is utilized to unscramble it. Hashing is frequently seen as a form of one-way encryption as you cannot go back from a hash to work out the first string; you can only go forward. In our proposed system, we read a fingerprint and then analyze the fingerprint and use base-64 hash to convert it into ciphertext. To avoid storing this ciphertext, we have treated this ciphertext as a password in terms of storing it in the database. Thus, we adopt an extra layer of security to hash the ciphertext using SHA-1 to satisfy the privacy of the user's fingerprint.

We utilize the salts technique to further thwart any rainbow table attack. Salts are short random sets of characters that are appended to the ends of a user password before they are hashed. Salts are automatically added after the user provides the fingerprint. Salts are generally stored in plaintext along with a hashed output, so the system knows which salt to use in regard to verifying a reset password.

Algorithm 1 Sign up using Fingerprint

```
Read User Email
Read Mobile number
Read Password
while not valid password do
    Read password
end while
password hashing(password)
send OTP to mobile
Read sent OTP from user
counter = 0
while is note valid OTP or counter < 3 do
    ** reenter the OTP message **
    Read sent OTP from user
    counter = counter + 1
end while
Read Fingerprint1 as Binary array
match rate = 0
while match rate < 75 do
    Read Fingerprint2 as binary array
    calculate match(Fingerprint1, Fingerprint2)
end while
match rate = 0
while match rate < 75 do
    Read Fingerprint3 as binary array
    calculate match(Fingerprint1, Fingerprint3)
end while
Encoding Fingerprint1 to Base64
Hashing Fingerprint1 SHA-1
Saver(email, mobile, password, Fingerprint)
```

VI. Experiment Results

We evaluate a simulated application of our proposed system so that we can have feedback about it while testing the usability of the system. The simulated application has been tested by eight participants. The average time to reset the password

Algorithm 2 Reset Password using Fingerprint

```
Read User Email
get mobile from storage
send OTP to registered mobile
Read OTP
counter = 0
while is not valid OTP or counter < 3 do
    ** reenter the OTP message **
    Read sent OTP from user
    counter = counter + 1
end while
get fingerprint from storage
is Match = False
while Match = False do
    Read Fingerprint as binary
    Encode Base64(user Fingerprint)
    Hashing Fingerprint1 SHA-1
    Match(user Fingerprint, Fingerprint from storage)
end while
Read New Password from user
hashing new password(new password)
update(password, new password)
```

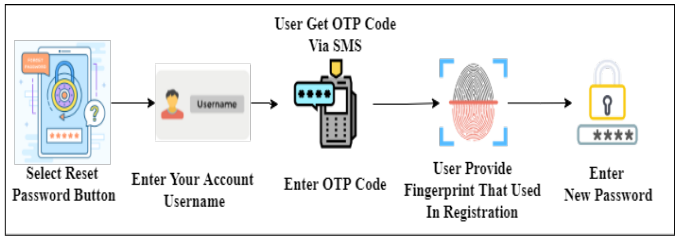


Fig. 4. Reset Password Process.

using the proposed system is 22 seconds that satisfies the user's experience without compromising the security, unlike the classical resetting system, which takes much time due to the challenge response system via email or other system. The evaluation gives us positive results on a lot of usability concerns, as presented in Fig. 5. These concerns included the time spent on the reset password process and the complicated process. 75% of the participants agreed on how easy and fast the password reset process is, while 100% agreed on the simplicity of the password reset process. Table II illustrates samples of participant's feedback.

TABLE II. SAMPLES OF PARTICIPANT FEEDBACK

Questions	Samples of Participant's Feedback
Differences between the classic resetting password and the proposed system.	'Many steps were shortened in the proposed system.'
	"Resetting the password in the proposed system is much easier than the available methods found in most of systems.'
	'In the proposed system, there is no need to leave the system, like email or SMS, and then through a link to do reset password.'
Is the proposed method tedious to reset password	'The proposed method is easy and not tedious'.
	Tow participants show concerns about saving the fingerprint template, whether or not it is safe.
Time	"It took short time".
complexity	"There is no complication".

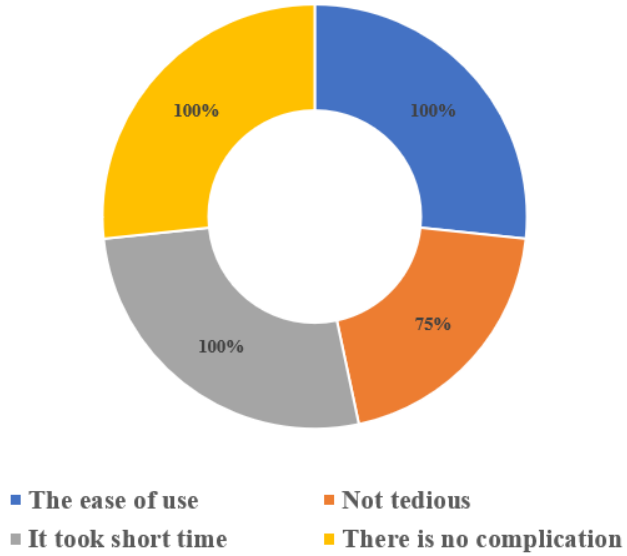


Fig. 5. Satisfaction Level of the Proposed System.

VII. CONCLUSION

Forgetting the password is a problem that exists and continues, it is part of human nature. The traditional method of resetting the password relies on increasing complexity, such as secret questions, which may affect the user experience. We propose a simulation of user authentication and the experiment includes a password-reset process. We use a fingerprint reader to emulate the mobile fingerprint sensor. Based on the simulation and evaluation results, the proposed system has several advantages over the challenge-responding system. The proposed system meets the security needs and at the same time provides usability. The user experience has a very large impact on usability. In the proposed system, to avoid direct saving of the template for the user's fingerprint, we encrypt the fingerprint and then use the Salt algorithm to be immune from decryption using available tools such as john the ripper and rainbow table. The new method of resetting the password gave us positive results in terms of usability and security. In the future, we are going to investigate the impact of biometric features in large scale systems and domains.

REFERENCES

[1] M. Theofanos, S. Garfinkel, and Y.-Y. Choong, "Secure and usable enterprise authentication: Lessons from the field," *IEEE Security & Privacy*, vol. 14, no. 5, pp. 14–21, 2016.

[2] S. Alotaibi, S. Furnell, and N. Clarke, "Transparent authentication systems for mobile device security: A review," in *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*. IEEE, 2015, pp. 406–413.

[3] H. Chen and G. Dong, "Fingerprint image enhancement by diffusion processes," in *2006 International Conference on Image Processing*. IEEE, 2006, pp. 297–300.

[4] J. Nielsen, "Usability 101: Introduction to usability (2012)," *URL: http://www. nngroup. com/articles/usability-101-introduction-to-usability/[Accessed November 2016]*, vol. 9, p. 35, 2012.

[5] O. Kulyk, S. Neumann, J. Budurushi, and M. Volkamer, "Nothing comes for free: How much usability can you sacrifice for security?" *IEEE Security & Privacy*, vol. 15, no. 3, pp. 24–29, 2017.

[6] R. Reid and J. Van Niekerk, "From information security to cyber security cultures," in *2014 Information Security for South Africa*. IEEE, 2014, pp. 1–7.

[7] S. Nørby, "Why forget? on the adaptive value of memory loss," *Perspectives on Psychological Science*, vol. 10, no. 5, pp. 551–578, 2015.

[8] L. F. Cranor and N. Buchler, "Better together: Usability and security go hand in hand," *IEEE Security & Privacy*, vol. 12, no. 6, pp. 89–93, 2014.

[9] S. M. R. S. Beheshti and P. Liatsis, "Captcha usability and performance, how to measure the usability level of human interactive applications quantitatively and qualitatively?" in *2015 International Conference on Developments of E-Systems Engineering (DeSE)*. IEEE, 2015, pp. 131–136.

[10] A. M. Eljetlawi, "Graphical password: Existing recognition base graphical password usability," in *INC2010: 6th international conference on networked computing*. IEEE, 2010, pp. 1–5.

[11] S. Pankanti, A. Jain, and L. Hong, "Biometrics: Promising frontiers for emerging identification market," *Comm. ACM*, pp. 91–98, 2000.

[12] J. ANDRESS, "Chapter 2–identification and authentication," *The Basics of Information Security (Second Edition). Se cond Edition. Boston: Syngress*, pp. 69–88, 2014.

[13] S. S. Ali, V. S. Baghel, I. I. Ganapathi, S. Prakash, S. Vu, and N. Werghi, "A novel technique for fingerprint based secure user authentication," *IEEE Transactions on Emerging Topics in Computing*, 2021.

[14] F. E. De Guzman, B. D. Gerardo, and R. P. Medina, "Implementation of enhanced secure hash algorithm towards a secured web portal," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*. IEEE, 2019, pp. 189–192.

Computational Analysis based on Advanced Correlation Automatic Detection Technology in BDD-FFS System

Xiao Zheng[1]
School of Computer Science and Technology
Shandong University of Technology
Xincun West Road, Zibo, Shandong,255090,China

Muhammad Tahir*[2]
School of Software Technology
Dalian University of Technology
Tuqiang Street, Dalian, 610024, China

Mingchu Li[3]
School of Software Technology
Dalian University of Technology
Tuqiang Street, Dalian, 610024, China

Shaoqing Wang[4]
School of Computer Science and Technology
Shandong University of Technology
Xincun West Road, Zibo, Shandong,255090,China

***Abstract*—Big Data-Driven Fabric Future Systems (BDD-FFS) is currently attracting widespread attention in the healthcare research community. Medical devices rely primarily on the intelligent Internet to gather important health-related information. According to this, we provide patients with deeply supportive data to help them through their recovery. However, due to the large number of medical devices, the address of the device can be modified by intruders, which can be life-threatening for serious patients (such as tumor patients). A large number of abnormal cells in the brain can lead to brain tumors, which harm brain tissue and can be life-threatening. Recognition of brain tumors at the beginning of the process is significant for their detection, prediction and therapy. The traditional approach for detecting is for a human to perform a biopsy and examine CT scans or magnetic resonance imaging (MRI), which is cumbersome,unrealistic for great amounts of resource, and requests the radiologist to make inferential computations. A variety of automation schemes have been designed to address these challenges. However, there is an urgent need to develop a technology that will detect brain tumors with remarkable accuracy in a much shorter time. In addition, the selection of feature sets for prediction is crucial to realize significant accuracy. This work utilizes an associative action learner with an advanced feature group, Partial Tree (PART-T), to detect brain tumor recognition grades. The model presented was compared with existing methods through 10-fold cross-validation. Experimental results show that partial trees with advanced feature sets are superior to existing techniques in terms of performance indicators used for evaluation, such as accuracy, recall rate and F-measure.**

Keywords*—Big data-driven fabric future systems (BDD-FFS); magnetic resonance imaging (MRI); partial tree

I. Introduction

Reliable and efficient collection and communication of observations from physical information systems that support the Internet of Things (IoT), for example sensors laid at faraway locations, to dominate centers are the current issues facing data detection in big data areas [1-3]. With the heavy use of devices in the healthcare trade, low-power communication equipment, and restricted reliability availability, there will be a variety of security threats. The most serious life threat facing the healthcare industry is brain tumors, which have a survival rate of less than 35 percent. With device management in the Internet of Things (IoT) [4-6], more specifically, in the Medical Internet of Things based on big data-driven future fabric (BDD-FFS) [7-8] systems, patient data can be hacked via botnets. As the result of, the security of BDD-FFS facilities is essential [9]. The disordered expansion of tissue in the brain was called a brain tumor, and it can lead to main (benign) or minor (malignant) tumors. A main tumor is a non cancer condition that does not propagate from one piece of the brain to another, yet a minor tumor is a cancer condition that maybe spread to other sections of the brain. In benign or malignant cases, the skull can compress and expand, causing damage to the brain and potentially life-threatening problems [10-11]. Therefore, accurate early prediction of brain tumors is crucial for their detection, prediction and therapy, which can only be ensured by exploiting safety schemes on the BDD-FFS device.

Brain tumors are often detected via biopsy, MRI or CT scan. In a biopsy test, a pathologist removes a few amount of material and checks it in line with a microscope to identify if there are signs of a tumor. While biopsy can accurately detect the presence of abnormalities, it can be unlucky for the patient. Second, the surgeon must be familiar with the accurate position and scale of the tumor when performing surgical tests.Therefore, MRI or CT scan is very important. One of the biggest merits of MRI compared with CT scan is that it is radiation-free, so it is healthy for human health. In addition, MRI can accurately detect tumors. However, extensive human MRI is a complex and unrealistic task, which relies on the technical awareness of doctors and technicians. In addition, a small number of radiologists will lead to higher cost and labor intensity of MRI analysis. Studies of approximations also suggest that radiologists filter out 15 to 25 percent of tumor treatments during screening [12]. Therefore, automatic recognition of brain tumors by MRI is a big change in medicine. Automatic recognition will allow doctors to predict the disordered growth of cells and tissues in the brain and

help repair early abnormalities. Many algorithms have been developed for calculating automatic recognition of brain tumors from MRI.

Although much of the work described above has validated various algorithms using computation to detect brain tumors, and some other experimental techniques have been executed to investigate brain tumors, the raised approach is known to be more accurate and precise. The method designed in this work has not been used to explore brain tumors in the above results. The complexity of partial trees is superior to all previous schemes. Partial trees showed better medical outcomes in terms of baseline and advanced features compared to other previous work.

The major contributions of the paper are the following:

(1) Aiming at the problem that traditional methods of brain tumor detection are cumbersome, unpractical and require radiologists reasoning time, a new technology is developed that can detect brain tumors with high accuracy in a shorter time.

(2) Since the selection of feature sets for prediction is crucial for real apparent accuracy. This work proposes an associative action trainer with a high-level feature union-PART tree (PARTT), to investigate the grade of brain tumor recognition.

(3) The results show that the partial trees with advanced feature sets are superior to the existing techniques in performance indicators such as accuracy, recall rate and F-measure.

The architecture of the rest of the papers is the following: Section II depicts the connect done in concerned areas, Section III portrays the materials and schemes involving the data union structure and raised mould, Section IV describes the outcomes and argument, and Section V supplies an survey of the paper and proposes some prospect discuss orientations.

II. Related Work

In literature [13], a computer assistance completely automated skill is exploited to detect glioma from multi-modle MRI images segment the tumor area from whole image. To diagnose tumors from brain MRI, a classifier called Naive Bayes is introduced. After detection, k-mean clustering and boundary prediction were served as collect brain tumor regions. The accuracy rate is more than 80%. In literature [14], large feature extraction proposals, in other sayings histogram of oriented gradients and gray rank concurrence matrix are applied to describe the graphicses. a segmentation scheme on account of color and edge detection was raised to investigate brain tumor regions. Budati aims at the problem that MRI is used as a result of low ionization and radiation in various medical imaging technologies, while manual detection takes a lot of trouble. Therefore, a machine learning technology is introduced to achieve the classification, recognition and detection effect of tumor or non-tumor areas in view of brain MRI dataset [15].

Automatic brain tumor detection, which graphics is segmented and classification is executed on brain MRI graphics applying genetic sequence that is meta-heuristic optimization scheme and support vector machine is proposed in [16]. The approach uses tumor attributes in genetic algorithms support vector machines feature extraction sociology. The tumor data classification is close to the user's views that are in view of dynamic style. The best Fuzzy in the light of Bayesian classification is a qualified method that has been raised to more the classification accurateness, only in case of the richness of details on these terraces allow them available for use as source data, in implantations depend on tumor cancer research fields [17]. Uma has a large number of violations in a specific website, and it is hard to analyze the data of the website. A new feature extraction method based on dependency parsing and sentiment dictionary is proposed. This feature is used in conjunction with dictionary-based features to classify specific data [18].

Pries et al. [19] concentrates on analysis of those research which comprise segmentation, detection and classification of brain tumors. The general process for a scheme which designed to classify brain tumors on FMRI or MRI scans is: Preprocessing the graphics such as though eliminating the noise, then segmenting the image, which generates the area which might be a brain tumor, and eventually classifying features like intensity, shape and texture of the area. The existing ML methods with regard to brain tumor detection have already been built. However, the above methods, even if emerging well outcomes, are not also employed. Therefore, all spatial pixels need to be converted into multi-directional ones.The application of Gabor transform for spatial to multidirectional image conversion is reflected in [20]. Gabor transform was applied to convert the noise filtered image into multi-dimension brain image. Mixed features such as GLCM, grayscale co-occurrence matrix and LDP, local derivative mode statistics and texture properties were calculated from the transformed brain images. Lu has developed a new computer-aided diagnostic system, called PBTNet, to validate the detection of primary brain tumors in MRI images because of the large differences between observers in the interpretation of MRI images [21].

Hazra was drawn in the detection and localization of tumor districts in the brain exploiting schemes raised by MRI images of patients [22]. The design scheme involves three steps: preprocessing, edge detection and segmentation. The pre-processing process relates converting the primitive graphics to grayscale image and eliminating the noise when it exists or sneaks in. This is accompanied with edge detection utilizing Sobel, Prewitt and Canny strategies with graphics enhancement technologies. Segmentation is then utilized so that the regions influenced via the tumor are obviously evident in the MRI image. Finally, K-means scheme is exploited to cluster images. Considerations on the correctness of tumor recognition and realistic placement of MRI images are given in literature [23]. A method for detecting brain tumors by easily utilizing magnetic resonance imaging (MRI) data was exploited in the research. Improve the quality of picture frame, convenient for patients to carry out tumor treatment and diagnosis. The proposed protocol improves MR image quality and the detection of brain tumors, making it easier for doctors to diagnose tumors. Due to the combination of automated image segmentation technology and automated and efficient brain tumor detection technology implemented on positron emission tomography (PET) images, Hagargi developed the operation and technology to detect brain tumors from PET images using artificial neural network (ANN). The network applies most artificial intelligence to the classification and

TABLE I. OUTLINE THE FRAMEWORK AND FEATURE SET DESCRIBED

Method	Segmentation	Features	Reference
Thresholding	Threshold	Area Set	Das et al. [8]
K-mean Clustering	Canny edge detection	Edge Det	Jos et al. [17]
Genetic Algorithm	Threshold	Region	Halder.[14]
Watersheding	Sobel Value	Metabolic threshold	Mus et al.[16]
MultiLayer Perceptron	Threshold	Fuzzy Algorithm	Sharma et al.[1
OTSU	Threshold	Intensity Gray	Singh et al.[15]
CART	EM Algorithm	K segmentation	Bh and Ch [11]
SVM	Threshold	Tumor area	Sing et al.[18]

recognition of biomedical images [24].

Tahir [25] studied an image filtering and grayscale segmentation method for feature extraction. The extracted feature union are transmitted to a deep neural network to identify tumor areas. A strategy combining threshold segmentation, feature extraction and filtering procedure is designed in literature [26]. Solidity, area, and bounding box are the features used for classification. In [2], a hybrid kernel-based support vector machine (HKSVM) was designed to identify brain tumors from MRI. At first, anisotropic filters are applied to images to isolate them from noise, and morphological operators are used to complete image segmentation. Isolate tumor areas using area prop algorithms. Feature vectors were extracted from isolated tumor regions according to GLCM and strength-based histogram. In the end, the extracted feature union was transmitted to HKSVM for tumor classification. The above schemes mainly emphasis on MRI segmentation, which is used to detect brain tumors.

In addition, the above method focuses on the binary classification of brain tumors as normal and abnormal, and does not involve level 1 (i.e., meningioma, metastasis), level 2 (i.e., low-level glioma), level 3 (i.e., glioma), and level 4 (i.e., glioma) in the light of the World Health Organization. However, in the above discuss, this paper adopts a new computational strategy, an association rule learner, Partial Tree, which makes use of some advanced feature sets that have not been used before to classify brain tumors into multiple categories. The set of advanced features exploited in the research are cell count, angle, density, perimeter, and center of mass, which are described in subsection 3. As can be seen from Table I, this feature group has not been used for tumor detection and judgment in previous studies. However, experimental results show that, compared with currently approaches (namely, Bayesian networks, random trees, Rep trees, random forests and Naive Bayes), the use of advanced feature set partial trees performs better in terms of time complexity and accuracy.

III. SYSTEM MODEL

A. Data Collection

The MRI was downloaded from various online sites. A total of 70 MRI samples were collected, including 32 samples of grade I, 10 samples of grade II, 8 samples of grade III and 20 samples of grade IV. Above all, the picked MRI was preprocessed to extract the feature group, and then the extracted feature information was delivered to the partial tree system to gradually classify brain tumors. The flow chart of the proposed approach is shown in **Fig.1**.

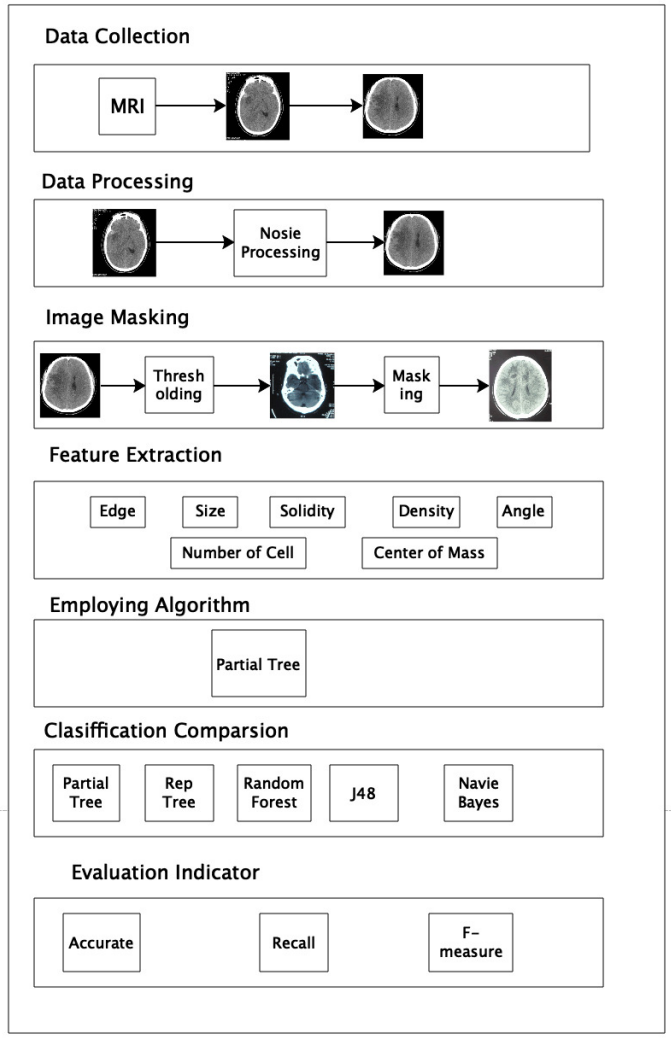


Fig. 1. Proposed Method Framework Diagram.

B. Feature Extraction

For the nature of feature extraction, some pre-processing processes are used for MRI, namely threshold segmentation and masking. According to the threshold segmentation, pixels are separated into different classification in view of their gray levels. The classification is regulated though the intensity value called the threshold. Pixels with larger intensity values fall into one classification, and pixels with smaller intensity values fall into another classification. It can be seen from the above literature that threshold is a common scheme that can successfully segment images. Therefore, the image can be segmented by threshold technology. In addition, the above data has not found the use of masking technology to identify brain tumors. The masking technology is introduced here as the preprocessing operation for detecting brain tumors, since in the case of the masking technology is appropriate for the image following the threshold, it will help to extract the features effectively.

Masking technology extracts features more effectively than other image processing schemes such as edge detection, mo-

tion detection and noise decrease, and is an effective performance of image processing. It can effectively measure the regional features and organizational structure in the image. Construct a duplication of the primary image and perform different AND, OR operations to meet its requirements. However, first use threshold segmentation on MRI, then construct a duplication of the image to employ the masking technique, and then use AND function to extract the required region. At the last, extract the features of the desired region. The feature set involves the number of cells, angular position, area, solidity, density, size, center of mass, perimeter, and so on. All of the above features are described the following.

1) Cell Number: Cell count refers to the entire amount of cells in the tumor area extracted. Use these factors to calculate the cell count using the following formula [27].

(1) The amount of small squares for counting according to the number of cells

(2) Count the number of large squares (tumor squares)

(3) Cell solution dilution

$$\text{The amount of cells} = \frac{\text{Number of large squares} \times \text{Cell solution dilution}}{\text{Number of small squares}}$$

2) Angle Position: The angle position shows details about the direction of tumor expansion in the skull. Tumors grow in a proportion of the brain that can travel vertically or horizontally through the brain at an angle.

3) Area: The size of the proportion shows particular details regarding the spatial spread of the tumor, namely, how much space the tumor can consume, in which S defines the region of the target (tumor) in the graphics, and x indicates the pixel value bigger than 1 and reaching N items in the region.

$$\sum_{x=1}^{N} Sx \tag{1}$$

4) Solidity: Solidity represents the density of the tumor and can be used to measure the number and size of depressions in the target boundary. The proportion of the brain tumor in the graphics target partitioned into the region of its convex hull is the solidity of the target graphics, which is expressed via formula (3)

$$Solidity = \frac{\text{Area Size}}{\text{Brain Convexhull}} \tag{2}$$

5) Density: Density is the key property applied to separate interest in an image region, which provides important information about image density.

6) Size: The size is considering the height and width of the objective image, which is built via multiplying though the number of horizontal and vertical pixels, as depicted in the Eq (4).

$$Size = \text{Horizontal pixels} \times \text{vertical pixels} \tag{3}$$

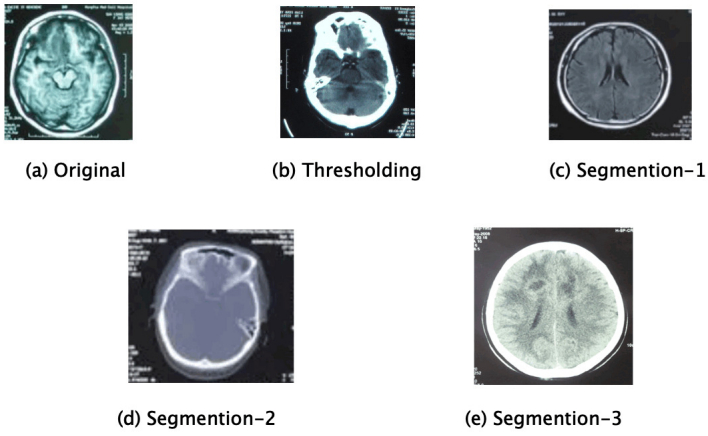


Fig. 2. Results from the Previous Process of Feature Extraction.

7) Center of Mass: The average value of each dimension in the target graphics is called the centroid, which is the median value of horizontal coordinate (x) and vertical coordinate (y) in the target graphics . The centroid is set by Eq. (5), which x, y and C respectively indicate the ordinate, abscissa and centroid of the object in the target image.

$$\text{Center of mass} = C \times \text{mean}(xory) \tag{4}$$

8) Perimeter: The perimeter refers to the sum of all external pixel boundaries in the target graphics. The most feasible approach to estimate the circumference of a brain tumor is to calculate the entire amount of edge pixels in the target graphics. Eq. (6) explains the representation of the perimeter.

$$\text{Perimeter} = \sum_{x,y} C(x, y) \tag{5}$$

which x and y define horizontal and vertical pixel values, while C represents the perimeter of the target in the target graphics.

The feature extraction procedure is shown in **Fig.2**. After feature extraction, the feature vector is transferred to the association rule learner, namely partial tree. The results demonstrate that the performance of partial tree is better than other algorithms, i.e., Random Trees, Rep Trees, Random Forests, and Naive Bayes are as above high-level feature collections on accuracy and time complexity.

IV. Partial Tree

Association rule learning tool has the ability to predict effectively. Therefore, this paper applies association rule mining model, that is, partial trees are used for grade detection of brain tumors. Partial tree is a mould designed by [28-29], which combines the merits of C4.5 and Ripper [30], and is used to yield a group of rules for efficient and accurate prediction. It takes advantage of Ripper's dial-and-conquer nature and

integrates it with C4.5 to prevent global optimization. C4.5 first builds an unpruned decision tree and transforms it into a rule union, and then simplifies rules by using a rule ordering strategy for each rule isolation. Finally, rules are set aside from the rule union to prevent global optimization until the rule set error is reduced. Ripper applies a divide-and-conquer approach to the set of rules. Only one rule can be introduced at a time, and the entities protected by this rule will be removed from the training sample. The rule generation process lasts up to the last entity of the training union. By combining the top features of C4.5 partial trees, a partial decision tree is initially constructed for the entities supplied in the dataset. Convert the leaves in the constructed tree with the largest coverage to rules, and then process the constructed partial tree to prevent global optimization. Entities protected by the yielded rule union are yet removed, and this process lasts until the final remaining entity in the dataset.

At first, some of the intelligibility of trees stimulated the use of them in hierarchical brain tumor detection in this paper. Second, it avoids global optimization, resulting in more noteworthy function in smaller time. In addition, it makes use of C4.5, which is a tree-based rule creation method, and from the existing research, tree-based method has the latent ability of identification. Therefore, a hybrid approach as a partial tree may perform better at extrapolating predictions. However, in this work, partial trees were found to have significant recognition of brain tumors and replace other existing methods such as Bayesian networks, random trees, Rep trees, random forests and naive Bayes in the light of time complexity and correctness.

V. Performance Evaluation

To verify the effectiveness of the algorithm, the following performance indicators are defined: accuracy (i.e., Acc), recall rate (i.e., Rec) and F-measure (i.e., F-mea). We use mathematical expressions (7), (8) and (9) respectively to describe the equations of these indicators.

$$Acc = \frac{TP}{TP + FP} \tag{6}$$

$$Rec = \frac{TP}{TP + FN} \tag{7}$$

$$F - mea = 2 \times \frac{Pre \times Rec}{Pre + Rec} \tag{8}$$

where TP represents true positive rate and is classified as brain tumor, FP is false positive rate, for instance, non-brain tumor is classified as brain tumor, FN is false negative rate, namely, brain tumor is classified as non-brain tumor. Consequently accuracy is the ratio between familiar instances of brain tumors that are correctly classified and all instances of brain tumors that are classified. Recall rate is the ratio between the amount of brain tumor instances that have been properly assorted and the whole amount of known brain tumor cases. The F-measure is the harmonic average of accuracy and recall rate.

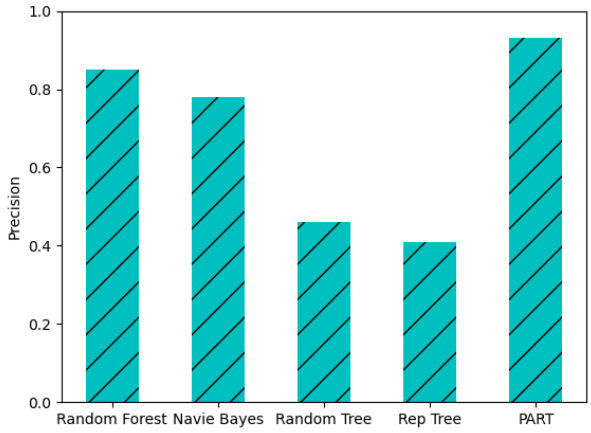


Fig. 3. Comparison of Accuracy between Partial Tree and Other Schemes.

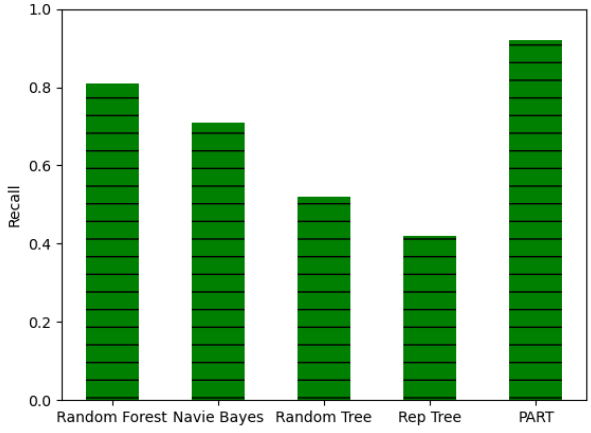


Fig. 4. Comparison of Recall with Partial Tree and Other Schemes.

A. Analysis of Experimental Results

To verify the efficiency of partial trees, it is compared with other competitive models such as random tree, Rep tree, random forest and naive Bayes. The performance comparison in accuracy, recall rate and F-measurement is shown in **Fig.3-5**. Compared with other technologies, partial tree will produce better results according to accuracy, recall rate and F-measure.

1) Performance Robustness: To verify the robustness of partial Tree performance, this paper compares it with the true positive (TP) rates of random Tree, Rep Tree, random forest and naive Bayes under different thresholds. The thresholds used are $t = 0.25$ to 0.95. **Fig.6** shows that the performance of some trees is more robust than that of other schemes because the TP rate is close to 1 while the TP rate of other methods is less than 0.85.

2) Computation Performance Cost: To verify the function of partial tree in computing cost, it is analyzed using mathematical and experimental algorithms respectively. The

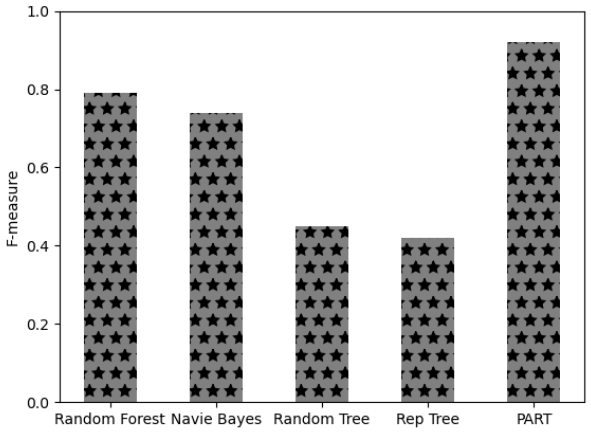


Fig. 5. Comparison of F-Measure with Partial Tree and Other Schemes.

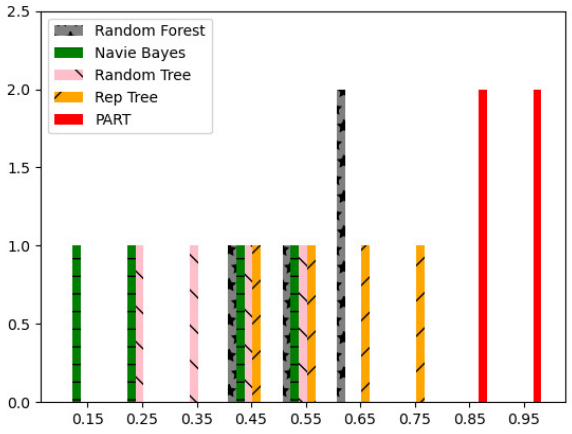


Fig. 6. Robustness Comparison of Algorithms under Different Thresholds.

TABLE II. ANALYSIS OF EXPERIMENTAL RESULTS

Algorithm	Correct Instances	Incorrectly Instances	Time (seconds)
Partial Tree	60	5	0.03
Rep Tree	30	35	1
Random Tree	32	27	1
Random Forest	50	15	0.3
Naive Bayes	48	18	0.2

TABLE III. THEORETICAL ANALYSIS OF DIFFERENT ALGORITHMS

Algorithm	Sample Size	Time Complexity (seconds)
Partial Tree	N+M	O($S \times NlogN$)
Rep Tree	N+M	O(N^2)
Random Tree	N+M	O(N^2)
Random Forest	N+M	O($M_{tree} \times N_{tree} \times NlogN$)
Naive Bayes	N+M	O($D \times N$)

mathematical performance analysis is shown in Table III, which N represents the amount of cases in the training and test set, where A denotes the amount of attributes, N_{tree} indicates the amount of trees built via the random forest, M_{tree} represents the amount of attributes sampled by every node of the tree, and D represents the dimension of features demanded by naive Bayes. According to Table III, for $\Omega(O)$, the time complexity of some trees is lower than that of Random Tree, Rep Tree and Random Forest algorithms. Compared with partial trees, naive Bayes is more effective in computation costs than partial trees, but less accurate than partial trees, as shown in Table III and Table II, which is very significant for correct diagnosis, prevention and therapy of diseases. Experimental analysis is denoted in Table II, which shows that it takes about 0.03 s to build a model for some trees. This is less time than other struggling approaches like random trees, Rep trees, and random forests.

Naive Bayes takes less time than partial trees. However, it can be concluded from Table III that the amount of cases of correct naive Bayes classification is smaller than the number of partial trees that may pose risks to the diagnosis, prevention and treatment of brain tumors.

Therefore, by studying an association rule learner with a different analysis called a partial tree, it can be concluded that some trees are superior to existing the most advanced technologies for instance Rep Tree, Random Tree, Random Forest and Naive Bayes in accordance with the correctness and computing expense. In addition, the set of high-ranking features described in the research, namely cell count, angular direction, density, centroid, and perimeter, still play a critical part in significantly making better the function of the brain tumor category-based model.

3) Impact of State-of-the-Art Features on Function: To verify the influence of state-of-the-art features presented in the research on brain tumor classification, the feature union was partitioned into two classes. The first class, called the benchmark feature, includes density, size, and area, whereas the second class, called the state-of-the-art class, includes cell count, Angle, density, perimeter, and center of mass. The all-around comparison of accuracy, recall rate and F-measure of

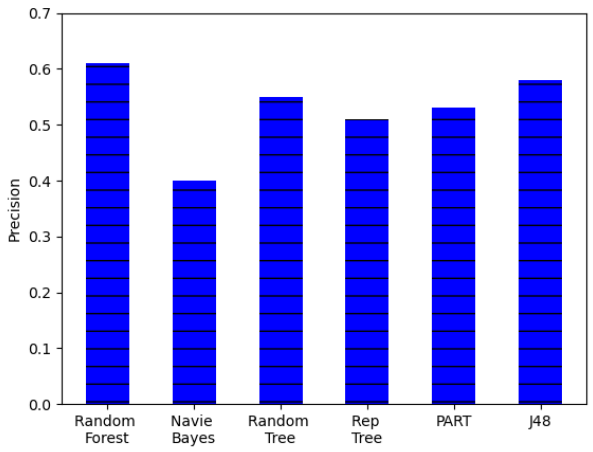


Fig. 7. Accuracy between Algorithms based on Baseline Features.

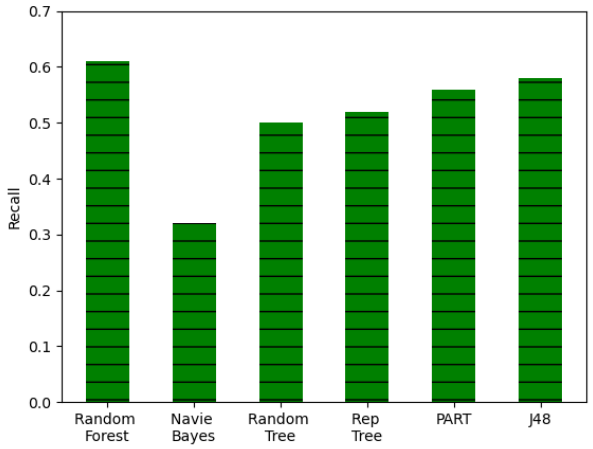


Fig. 8. Various Recall Algorithms based on Baseline Features.

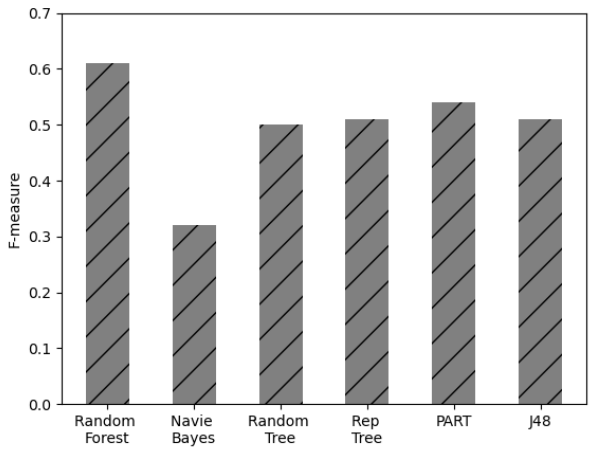


Fig. 9. F-Measure in Different Algorithms based on Baseline Feature.

Partial Tree, J48, Random Tree, Rep Tree, Random Forest and Naive Bayes is shown in **Fig. 7 to 9**, and baseline feature sets are used respectively. **Fig. 10 to 12** shows the comprehensive comparison of accuracy, recall rate and F-measure of Partial Tress, J48, Random Tree, Rep Tree, Random Forest and Naive Bayes, separately employing high-level function union features. It is important to note that the most state-of-the-art methods have baseline characteristics, accuracy, recall rates, and F-measures of less than 60%. For another, with the improvement of advanced feature sets, these methods increase significantly in accuracy, recall rate and F-measure. However, Partial Tress and J48 with 96% accuracy, recall and F-measure will show superior than other schemes. J48 has 60% F-Measure, but its performance is significantly improved through the advanced feature set, as shown in Fig. 10 to 12. Therefore, the function of distinct machine learning models should be better with advanced feature sets.

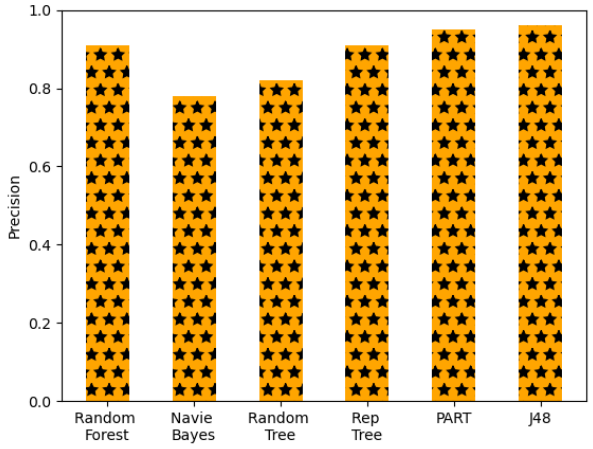


Fig. 10. Accuracy of Various Algorithms based on Advanced Feature Sets.

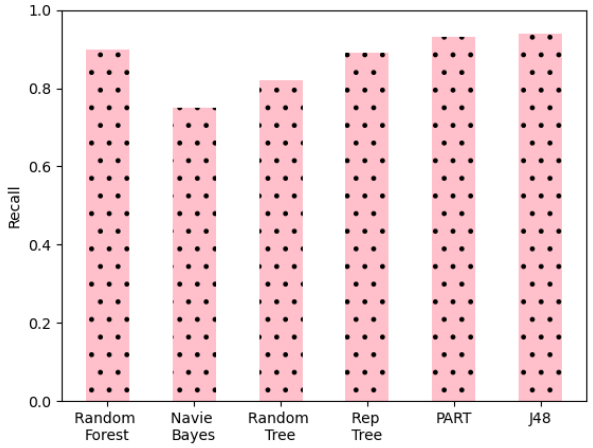


Fig. 11. Recall of Various Algorithms based on Advanced Feature Sets.

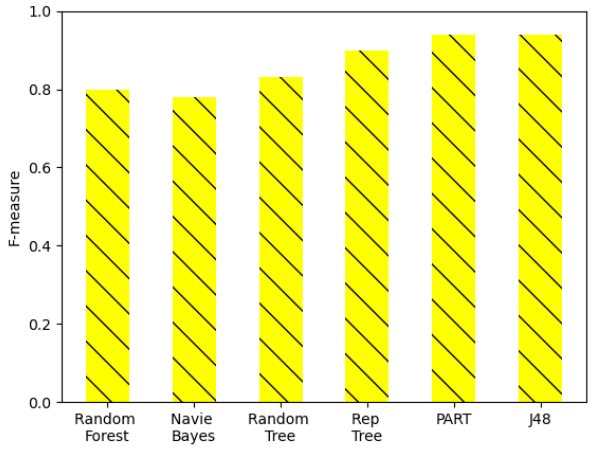


Fig. 12. F-Measure of Different Algorithms based on Advanced Feature Sets.

VI. Conclusion

With the emergence of the BDD_FFS system in the medical field, there have been many security risks in big data medical equipment. However, the most serious security threat facing the healthcare field is brain tumors. Brain tumors are irregular growths of cell tissue in the brain, which may lead to life-threatening. In a safe environment, the use of BDD_FFS-based system applications to detect early brain tumors is important to reduce mortality. The conventional techniques used to detect brain tumors are biopsy and MRI or CT scanning by human experts. Biopsies are very unbearable for patients, and a large number of MRI and CT scans is a complex task that is impractical for limited specialists.

Therefore, it is necessary to adopt safe and automatic technology to accurately detect brain tumors. In this work, firstly, a secure PART_T-based computational approach is employed to correctly identify the lesion grade of brain tumors. Secondly, this work introduces a high-ranking feature union not formerly involved for the proper recognition of brain tumors. Finally, The experimental results show that the designed PART_T technology with advanced feature group is superior to other subsistent technologies for instance Rep Tree, Random Tree, Random Forest and Naive Bayes in correctness and computational overhead.

The future work is as follows: expert automation technology system is very necessary to determine brain tumors in the early period, so that they can be more treated with drugs, so as to avoid a series of processes of surgical pain. Firstly, Applying techniques such as compounded machine learning models or neural networks to improve brain tumor diagnostic procedures is an inevitable need for active academic attention. Secondly,, some further modern functions can also be applied to make better medical effects. BDD_FFS combines active metadata, semantics, knowledge mapping, data virtualization, AI and other technologies to enable accurate, agile, and efficient matching between users and data to achieve optimization for specific medical scenarios. Finally, BDD_FFS is a technical architecture approach that addresses the complexity of data and metadata in an intelligent way and provides seamless access and sharing for all data consumers.

Acknowledgment

This work was supported in part by the National Nature Science Foundation of China under Grant no. 61572095, 61877007, 61802097, and in part by the Project of Qianjiang Talent under Grant no. QJD1802020. This work also was supported by Shandong Provincial Natural Science Foundation, China (ZR2020MF147).

References

[1] M. Faheem, G. Fizza, M. W. Ashraf, R. A. Butt, and V. C. Gungor. Data acquired by internet of things-enabled industrial multichannel wireless sensors networks for active monitoring and control in the smart grid industry 4.0. Data in Brief, 35(4):106854, 2021.

[2] JRR García, Martinetti A , Becker J , et al. Towards an Industry 4.0-Based Maintenance Approach in the Manufacturing Processes[M]. 2021.

[3] M. Faheem, M. Umar, R. A. Butt, B. Raza, M. A. Ngadi, and V. C. Gungor. Software defined communication framework for smart grid to meet energy demands in smart cities. In 2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG), 2019.

[4] B Dwa, B Dza, and C As. Energy management solutions in the internet of things applications: Technical analysis and new research directions. Cognitive Systems Research, 2021.

[5] L. Wang and Y. Dong. Stochastic neural network based data analysisrelated talent recruitment optimization via cdn server. Internet Technology Letters, 2020.

[6] M. S. Haghighi, M. Ebrahimi, S. Garg, and A. Jolfaei. Intelligent trustbased public-key management for iot by linking edge devices in a fog architecture. IEEE Internet of Things Journal, 8(16):12716-12723, 2021.

[7] N. Kesswani and S. Choudhary. A survey: Intrusion detection techniques for internet of things. International Journal of Information Security and Privacy, 13(1):86-105, 2019.

[8] Y. Tai, B. Gao, Q. Li, Z. Yu, and V. Chang. Trustworthy and intelligent covid diagnostic iomt through xr and deep learning-based clinic data access. IEEE Internet of Things Journal, PP(99):11, 2021.

[9] Sana Ullah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din, Atif Khan, and Yasir Faheem. An e-health care services framework for the detection and classification of breast cancer in breast cytology images as an iomt application. Future Generation Computer Systems, 98:286-296, 2019.

[10] Nitish and Amit Kumar Singh. Automatic detection classification and area calculation of brain tumour in mri using wavelet transform and svm classifier. International Journal of Intelligent Systems Technologies and Applications, 19(6):526-540, 2020.

[11] Ikram Ud Din, Mohsen Guizani, Joel J.P.C. Rodrigues, Suhaidi Hassan, and Valery V. Korotaev. Machine learning in the internet of things: Designed techniques for smart cities. Future Generation Computer Systems, 100:826-843, 2019.

[12] T. Kavitha, S. Hemalatha, and C. Subhashini. Mri image segmentation and detection in image processing for brain tumor. 2018.

[13] I. Zabir, S. Paul, M. A. Rayhan, T. Sarker, and C. Shahnaz. Automatic brain tumor detection and segmentation from multi-modal mri images based on region growing and level set evolution. In 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 2016.

[14] S. Chauhan, A. More, R. Uikey, P. Malviya, and A. Moghe. Brain tumor detection and classification in mri images using image and data mining. In International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE), 2017.

[15] Budati A K , Katta R B . An automated brain tumor detection and classification from MRI images using machine learning techniques with IoT[J]. Environment, Development and Sustainability, 2021:1-15.

[16] T. L. Narayana and T. S. Reddy. An efficient optimization technique to detect brain tumor from mri images. In International Conference on Smart Systems and Inventive Technology, 2018.

[17] M. Karthica and P. Sudarmani. Fbsc: An analyzing sentiments using fuzzy based bayesian classification. 2019.

[18] Uma M , Florence S M , Jesi V E , et al. Analysis of Ensemble Classification of Twitter Sentiments Using New Dependency Tree Based Approach[J]. International Journal on Artificial Intelligence Tools, 2022, 31(05).

[19] TP Pries, R. Jahan, and P. Suman. Review of brain tumor segmentation, detection and classification algorithms in fmri images. In 2018 International Conference on Computational and Characterization Techniques in Engineering Sciences (CCTES), 2018.

[20] V. Jeevanantham and G. Mohanbabu. Detection and diagnosis of brain tumors-framework using extreme machine learning and canfis classification algorithms. International Journal of Imaging Systems and Technology, 2020.

[21] Lu S , Satapathy S , Wang S , et al. PBTNet: A New Computer-Aided Diagnosis System for Detecting Primary Brain Tumors.[J]. Frontiers in cell and developmental biology, 2021, 9:765654.

[22] A. Hazra, A. Dey, S. K. Gupta, and M. A. Ansari. Brain tumor detection based on segmentation using matlab. In International Conference on Energy, Communication, Data Analytics and Soft Computing, Chennai(IN), 2019.

[23] F. Rehman, H. Panhwar, S. Rajpar, S. S. Shah, and S. Rabani. Brain tumor detection from mr images using image process techniques and tools in matlab software. Journal of Advanced Medical Sciences and Applied Technologies, 1(4):1-5, 2021.

[24] Hagargi P A . Brain Tumor Detection using ANN Algorithm[J]. 2021.

[25] Muhammad Naeem Tahir. Classification and characterization of brain tumor mri by using gray scaled segmentationand dnn. 2018.

[26] Thomas A. Roberts, Harpreet Hyare, Giulia Agliardi, Ben Hipwell, Angela d'Esposito, Andrada Ianus, James O. Breen-Norris, Rajiv Ramasawmy, Valerie Taylor, David Atkinson, Shonit Punwani, Mark F. Lythgoe, Bernard Siow, Sebastian Brandner, Jeremy Rees, Eleftheria Panagiotaki, Daniel C. Alexander, and Simon Walker-Samuel. Non-invasive diffusion magnetic resonance imaging of brain tumour cell size for the early detection of therapeutic response. Scientific Reports, 10(1):9223, 2020.

[27] Haris K , Valsan V , Pai A . Performance of multi-detector hybrid statistic in targeted compact binary coalescence search[J]. 2017.

[28] Jaworski M , Duda P , Rutkowski L . New Splitting Criteria for Decision Trees in Stationary Data Streams[J]. IEEE Transactions on Neural Networks Learning Systems, 2018, 29(6):2516-2529.

[29] Marongiu M , Pellizzoni A , Egron E , et al. Methods for detection and analysis of weak radio sources with single-dish radio telescopes[J]. Experimental Astronomy, 2020, 49(2).

[30] Gantenbein M , Wang L , Al-Jobory A A , et al. Quantum interference and heteroaromaticity of para- and meta-linked bridged biphenyl units in single molecular conductance measurements[J]. Scientific Reports, 2017, 7(1).

Image Enhancement Method based on an Improved Fuzzy C-Means Clustering

Libao Yang
Faculty of Science and Natural Resources,Universiti Malaysia Sabah, Kota Kinabalu,88400

Suzelawati Zenian
Faculty of Science and Natural Resources,Universiti Malaysia Sabah, Kota Kinabalu,88400

Rozaimi Zakaria
Faculty of Science and Natural Resources,Universiti Malaysia Sabah, Kota Kinabalu,88400

***Abstract*—Image enhancement is an important method in the process of image processing. This paper proposes an image enhancement method base on an improved fuzzy c-means clustering. The method consists of the following steps: firstly, proposed a fuzzy c-means clustering with a cooperation center(FCM-co). Secondly, using the FCM-co, divide the image pixels into different clusters and marked membership values to those clusters. Thirdly, modify the membership values. Finally, calculate the new pixel gray levels. This enhancement method can overcome the disadvantage of overexposure and better retain image details. Through the experiment, the test results show that the proposed enhancement method could achieve better performance.**

***Keywords*—Image enhancement; fuzzy clustering; fuzzy c-means clustering; membership; objective function**

I. Introduction

Image enhancement plays a significant role in digital image processing. Low contrast in digital images can result from many circumstances, including lack of sunlight or indoor lighting, and inadequacy of the device. There are many methods to enhance images. Histogram equalization (HE) is the simplest image enhancement method. It stretches the histogram of the image, based on the probability probability density function and cumulative distribution function values of the pixels, leading to enhancement in the contrast of the image [1], [2], [3], [4], [5]. The gamma correction-based method is an automatic transformation technique that improves the brightness of dimmed images via the gamma correction and probability distribution of luminance pixels, this method uses temporal information regarding the differences between each frame to reduce computational complexity [6], [7], [8]. Fuzzy sets can deal with some uncertain factors better than classical mathematics. Fuzzy technology is also increasingly used for image processing [9], [10], [11].

In 2000, H.D. Cheng et.al proposed a novel adaptive direct fuzzy contrast enhancement method based on the fuzzy entropy principle and fuzzy set theory [12]. In 2009, M. Hanmandlu et.al presented a new approach for the enhancement of color images using the fuzzy logic technique [13]. In 2011, G. Li et.al proposed an image enhancement operation that used the value of grey entropy in the neighborhood window as parameters to measure the level of the current pixel being edge point [14]. In 2012, K. Hasikin et.al presented a fuzzy gray scale enhancement technique for low contrast image [15]. In 2016, A. K. Gupta et.al presented a fuzzy based enhancement technique for low contrast gray scale image [16]. In 2017, V. Magudeeswaran et.al presented a Contrast limited fuzzy adaptive histogram equalization to improve the contrast of MRI Brain images [17]. In 2019, S. Zenian et.al implemented an intuitionistic fuzzy set and fuzzy set, respectively, in the fEEG image by using intensification operator in enhancing the contrast of the image [18], [19].

The fuzzy c-means (FCM) clustering algorithm was first introduced by Dunn [20] and later extended by Bezdek [21]. FCM clustering algorithm is often used to deal with data classification problems. In recent years, it has been applied to image processing [22]. The main idea of the algorithm is to divide the data set into different categories by calculating the difference between gray values and clustering center iteration, so as to optimize the criterion function for evaluating clustering performance.The algorithm is an iterative clustering method that produces an optimal partition by minimizing the weighted within group sum of the squared error objective function.

This paper proposed an image enhancement method base on an improved fuzzy c-means clustering (FCM-co). Compared with the traditional fuzzy C-means clustering, FCM-co has a cooperation center. The data used in cooperation center calculate is a cooperation matrix of the same size as the image. The cooperation matrix element value is the average of the gray values of the pixel at the corresponding image position and the pixels around it. This means that FCM-co also considers the image pixels' location information. In the clustering process, the cooperation center is always updated synchronously with the clustering center. After the FCM-co divides the image pixels into different clusters and marks pixels' membership value, we modify the pixels' membership value again and calculate the pixels' new gray levels. The paper's contribution is to propose a new clustering method(FCM-co) and a new function to modify the membership value. In the last section, the test results show that this paper proposed an enhancement method that could achieve better performance.

II. Methodology

A. Improve Fuzzy C-Means Clustering

Compared with traditional clustering methods, in order to better use the position information of pixels in the image, this paper proposes a fuzzy c-means clustering with a cooperation center (FCM-co). The FCM-co's objective function as follows:

$$J_m = \sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij}^{m} \left(\|x_j - v_i\|^2 + \alpha \left\|x_j^* - v_i^*\right\|^2 \right). \quad (1)$$

Image $I = \{x_{k,p}|k = 1, 2, 3, \cdots, m, p = 1, 2, 3, \cdots, n\}$, where $x_{k,p}$ is the gray scale of the pixel in row k and column p of the image. $I^* = \{x^*_{k,p}|x^*_{k,p} = mean(x_{k,p}$ and around $x_{k,p})\}$ is the cooperation matrix of Image I. In the Eq. (1), $X = \{x_j|x_j = x_{k,p}, j = (k-1)n+p$, 1 ≤ k ≤ m, 1 ≤ p ≤ n, $x_{k,p} \in I\}$, similarly, $X^* = \{x^*_j|x^*_j = x^*_{k,p}, j = (k-1)n+p$, 1 ≤ k ≤ m, 1 ≤ p ≤ n, $x^*_{k,p} \in I^*\}$, $N = mn$. c is the number of clusters. u_{ij} is the degree of membership of x_j and x^*_j in ith cluster, m is the weighting exponent on each fuzzy membership, v_i and v^*_i are the prototype of the center of cluster i, $\|x_j - v_i\|^2$ is a distance measure between object x_j and cluster center v_i, $\left\|x^*_j - v^*_i\right\|^2$ is a distance measure between object x^*_j and cluster center v^*_i. The parameter α is a constant. By definition, each point x_j satisfies the constraint that $\sum_i^c u_{ij} = 1$. The object function J_m can be obtained through an iterative as follows:

Step A : Initialize the membership values u_{ij}.

Step B : Calculate the v_i and v^*_i by

$$v_i = \frac{\sum_{j=1}^N u_{ij}^m x_j}{\sum_{j=1}^N u_{ij}^m}, \tag{2}$$

and

$$v^*_i = \frac{\sum_{j=1}^N u_{ij}^m x^*_j}{\sum_{j=1}^N u_{ij}^m}. \tag{3}$$

Step C : Update u_{ij}

$$u_{ij} = \sum_{k=1}^{c} \left(\frac{\|x_j - v_i\|^2 + \alpha \left\|x^*_j - v^*_i\right\|^2}{\|x_j - v_k\|^2 + \alpha \left\|x^*_j - v^*_k\right\|^2} \right)^{-\frac{1}{m-1}}. \tag{4}$$

Step D : Compute the value of the objective function $J_m^{(t)}$

$$J_m^{(t)} = \sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij}^m \left(\|x_j - v_i\|^2 + \alpha \left\|x^*_j - v^*_i\right\|^2 \right). \tag{5}$$

Step E : If $\left|J_m^{(t)} - J_m^{(t-1)}\right| < \epsilon$, then stop. Otherwise, $t = t + 1$, return to step B.

B. Modify Membership and Calculate New Gray Scale Level

After FCM-co marks the membership value for pixels, for further adjust pixels' the membership value, we propose the following adjustment function:

$$u^*_{ij}(x_j) = \begin{cases} u_{ij}(x_j), & x_j \le \hat{v}_{i=1,2,\cdots,c} \\ 1 + \frac{1-u_{ij}(x_j)}{2}, & x_j > \hat{v}_{i=1,2,\cdots,c} \end{cases}. \tag{6}$$

In the Eq.(6), $j = 1, 2, 3, \cdots, N$. $u^*_{ij}(x)$ is the modified membership of the jth pixel in the ith cluster. $\hat{v}_i$ is the cluster, $\hat{v}_i$=max(v_i, v^*_i).

For calculation of the new gray scale level of the pixel, the original image gray scale levels are updated and mapped to compute the enhanced image by given formulations:

$$y_j = \frac{1}{c} \sum_{i=1}^{c} u^*_{ij}(x_j) x_j, \quad j = 1, 2, \cdots, N. \tag{7}$$

Where y_j is the new gray scale level of the jth pixel.

C. Algorithm

This paper uses the algorithm to process the test images as follows:

Step 1: Initialize the parameters: m, c, α, and ϵ.

Step 2: Calculate the cluster centers and pixels' membership value using FCM-co, Eqs.(2)-(5).

Step 3: Modify the membership values using Eq.(6).

Step 4: Calculate the new pixels' gray scale level using Eq.(7).

III. Experimental Results and Analysis

In this section, when we use the algorithm in the experiment, set the the parameters $m = 2, c = 2, \alpha = 1$, and $\epsilon = 0.00001$.

A. Subjective Analysis

There show the effect of the proposed method on image enhancement (see Fig. 1). To analyze the performance, the proposed method is compared with methods in [23], [24], and [25] (see Fig. 2).

Fig. 1. Original and Result Images.(a) Original, (b) Enhanced by Proposed Method.

Fig. 1 shows that the proposed method can enhance the image. In Fig. 2(b), and Fig. 2(d), the layered highland, signage, door inside the courtyard wall, and house in the distance are not visible. It overexposed the image and lost some image details. Although Fig. 2(c) retains more image details, it also has the disadvantage of insufficient enhancement. The image details retained by Fig. 2(e) are similar to those of Fig. 2(c), such as layered highland, signage, door inside the courtyard wall, and house. But the image contrast of Fig.

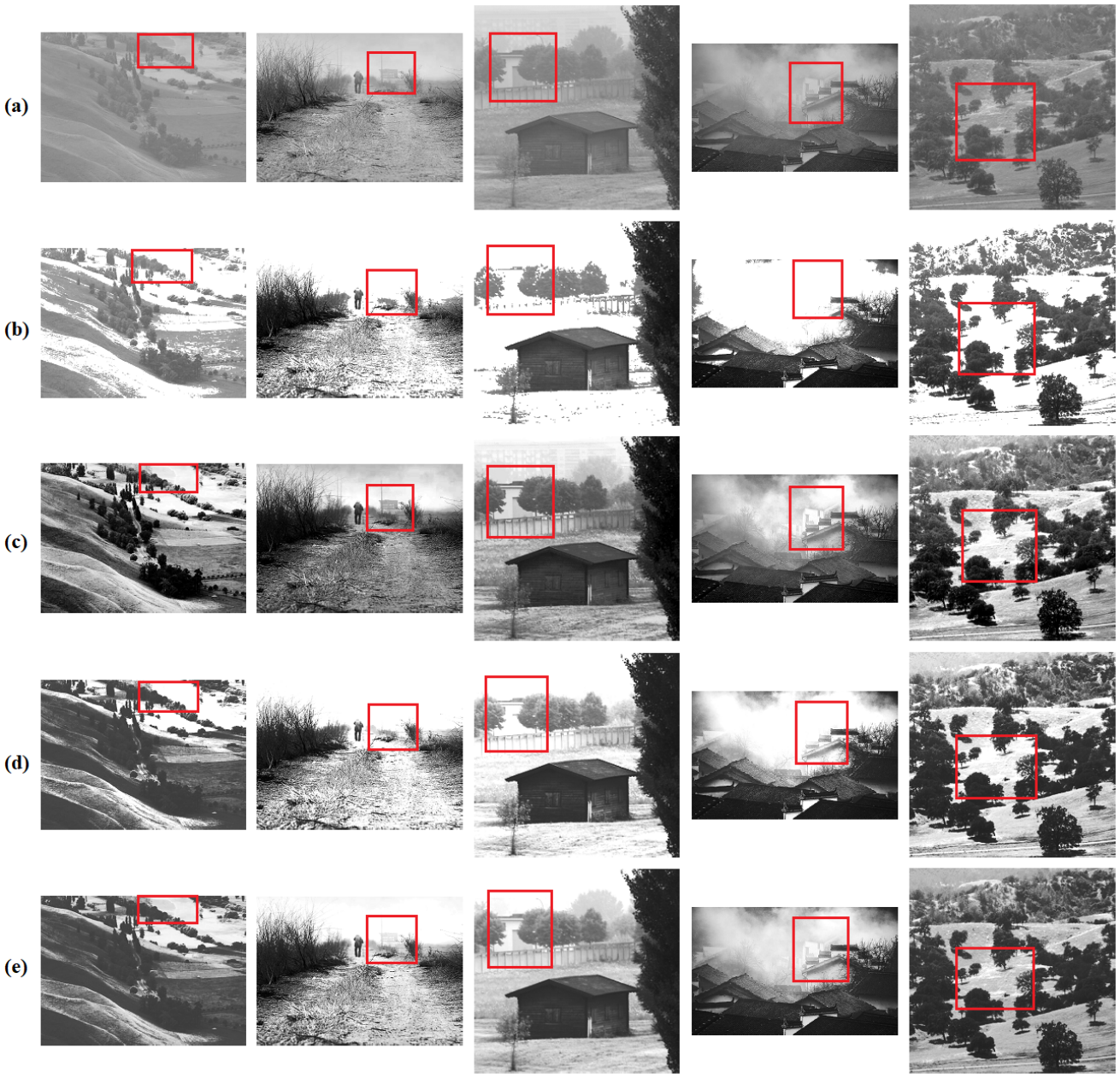


Fig. 2. Original and Result Images. (a) Original, (b), (c) and (d) are Enhanced by Methods in [23], [24], and [25], Respectively, (e) Enhanced by Proposed Method.

2(e) is higher than that of Fig. 2(c). Through visual contrast, Fig. 2(e) (processed by the proposed method) can increase the image contrast, be fully exposed, and also retain some obvious image details.

B. Objective Analysis

For image enhancement effect evaluation, this paper used algorithms include mean squared error(MSE), peak signal-noise ratio(PSNR), structural similarity(SSIM), average gradient(AG), Linear index of fuzziness(IOF) and entropy[26], [27]. The lower MSE(IOF) or the higher PSNR(SSIM, AG, entropy) indicates a better enhancement effect.

$$MSE = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (x_{i,j} - y_{i,j})^2. \tag{8}$$

$$PSNR = 10 \times \log_{10} \frac{(2^n - 1)^2}{MSE} dB. \tag{9}$$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}. \tag{10}$$

TABLE I. MEAN SQUARED ERROR (MSE) TEST RESULTS

Image \ Method	in [24]	in [23]	in [25]	Proposed
APC	3319.11	10067.11	3965.29	**1029.43**
Aerial	**1157.93**	2952.3	2401.57	1875.44
Aerial2	4724.8	1990.49	2600.87	**1691.43**
Airplane (U-2)	12512.85	9470.55	358.98	**197.05**
Airplane	6636.69	**1403.04**	2325.95	2059.08
Airplane2	6171.68	3240.9	5274.58	**1923.66**
Airport	3929.97	7638.73	1462.95	**802.09**
Car and APCs	2222.48	8356.03	3081.27	**875.94**
Car and APCs2	2604.8	7992.16	2987.44	**1015.41**
Chemical plant	1758.64	7479.43	2158.83	**852.43**
Clock	4692.85	1325	1391.97	**1190.81**
Couple	1553.43	6338.95	2794.77	**1008.98**
Fishing Boat	1305.72	6892.04	3759.02	**1097.29**
Male	1818.22	7950.11	2110.15	**656.78**
Moon surface	2552.64	6804.04	3166.42	**1280.56**
Stream and bridge	**741.44**	4422.23	1811.54	1151.65
Tank	2724.2	7717.86	3971.38	**1175.79**
Tank2	3148.4	9525.06	2327.63	**919.98**
Tank3	1638.34	6282.33	3529.53	**1298.46**
Truck and APCs	2098.07	8009.45	2368.24	**883.67**
Truck and APCs2	3165.58	9909.88	1786.73	**653.02**
Truck	2915.86	10696.75	2598.83	**857.2**

TABLE II. PEAK SIGNAL TO NOISE RATIO (PSNR) TEST RESULTS

Image \ Method	in [24]	in [23]	in [25]	Proposed
APC	12.9206	8.1018	12.1481	**18.0048**
Aerial	**17.4940**	13.4292	14.3258	15.3998
Aerial2	11.3870	15.1412	13.9796	**15.8483**
Airplane (U-2)	7.1572	8.3671	22.5802	**25.1850**
Airplane	9.9113	**16.6601**	14.4648	14.9941
Airplane2	10.2268	13.0241	10.9089	**15.2895**
Airport	12.1869	9.3006	16.4785	**19.0886**
Car and APCs	14.6624	8.9108	13.2435	**18.7061**
Car and APCs2	13.9731	9.1042	13.3778	**18.0644**
Chemical plant	15.6790	9.3921	14.7886	**18.8242**
Clock	11.4164	16.9086	16.6945	**17.3724**
Couple	16.2179	10.1106	13.6673	**18.0920**
Fishing Boat	16.9723	9.7473	12.3801	**17.7276**
Male	15.5343	9.1271	14.8877	**19.9566**
Moon surface	14.0609	9.8031	13.1251	**17.0568**
Stream and bridge	**19.4300**	11.6744	15.5503	17.5176
Tank	13.7784	9.2558	12.1414	**17.4275**
Tank2	13.1499	8.3421	14.4617	**18.4930**
Tank3	15.9868	10.1496	12.6536	**16.9965**
Truck and APCs	14.9126	9.0948	14.3865	**18.6679**
Truck and APCs2	13.1263	8.1701	15.6102	**19.9815**
Truck	13.4831	7.8383	13.9830	**18.8000**

$$AG = \frac{1}{(M-1)\times(N-1)} \sum_{i=1}^{M-1}\sum_{j=1}^{N-1} \sqrt{\frac{(x_{i,j}-x_{i+1,j})^2 + (x_{i,j}-x_{i,j+1})^2}{2}}. \tag{11}$$

$$IOF = \frac{2}{M \times N} \sum_{i=1}^{M}\sum_{j=1}^{N} min(u_{ij}^{'}, 1-u_{ij}^{'}). \tag{12}$$

In Eq. (9), $n = 8$ (the test image are 8bit image). In equation (10), μ_x is the mean of x, μ_y is the mean of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y, σ_{xy} is the covariance of x and y. c_1 and c_2 are constants[28]. In Eq. (12), $u_{ij}^{'}$ is membership value. For experimentation, we considered 22 images from Miscellaneous(MISC) dataset (http://sipi.usc.edu/database/dat abase.php?volume=misc).

As shown in Tables I, II, and III, except for test image

TABLE III. STRUCTURAL SIMILARITY (SSIM) TEST RESULTS

Image \ Method	in [24]	in [23]	in [25]	Proposed
APC	0.3585	0.4292	0.7818	**0.8963**
Aerial	**0.8055**	0.6303	0.6823	0.7630
Aerial2	0.6148	0.7503	0.7309	**0.7966**
Airplane (U-2)	0.1946	0.3386	0.8051	**0.8148**
Airplane	0.5667	**0.8589**	0.8027	0.8421
Airplane2	0.3827	0.6809	0.8653	**0.9411**
Airport	0.6645	0.5493	0.7504	**0.8119**
Car and APCs	0.5977	0.4896	0.7244	**0.8392**
Car and APCs2	0.5234	0.4578	0.6266	**0.7468**
Chemical plant	0.7740	0.4812	0.6999	**0.8096**
Clock	0.5808	0.8662	0.8690	**0.8916**
Couple	0.7042	0.5386	0.6924	**0.8006**
Fishing Boat	0.7017	0.5770	0.7870	**0.8917**
Male	0.8412	0.5518	0.8182	**0.8862**
Moon surface	0.5188	0.4498	0.5805	**0.7448**
Stream and bridge	**0.8851**	0.5854	0.7310	0.7992
Tank	0.4418	0.5170	0.6944	**0.8216**
Tank2	0.5034	0.3922	0.5727	**0.6836**
Tank3	0.6819	0.5882	0.7069	**0.8080**
Truck and APCs	0.6861	0.4794	0.6629	**0.7697**
Truck and APCs2	0.6718	0.4302	0.6987	**0.7985**
Truck	0.5795	0.4567	0.6905	**0.8044**

TABLE IV. AVERAGE GRADIENT (AG) TEST RESULTS

Image \ Method	no processed	in [23]	in [24]	in [25]	Proposed
APC	5.7959	11.4499	8.2862	23.6886	**30.4575**
Aerial	16.3869	30.7155	25.6089	**30.8756**	27.5800
Aerial2	12.3869	20.1639	19.4104	21.2859	**25.3468**
Airplane (U-2)	6.0065	5.5144	4.4803	32.7717	**40.3657**
Airplane	4.1840	6.5466	5.5092	6.2647	**8.5707**
Airplane2	3.3934	3.3481	4.8411	10.1661	**19.2798**
Airport	11.1810	19.7410	14.5196	**28.9451**	26.4079
Car and APCs	6.3992	13.9373	9.8210	16.9770	**17.2540**
Car and APCs2	6.5654	17.3508	11.6941	**20.7892**	20.2743
Chemical plant	12.1112	23.8362	17.8722	**27.8279**	21.8129
Clock	6.9641	9.2652	7.8369	9.4388	**9.9800**
Couple	8.2969	16.6550	12.4401	**18.7798**	16.4511
Fishing Boat	9.3853	15.4133	12.7929	15.6537	**20.3534**
Male	7.9166	11.6793	9.5091	**11.9637**	10.6810
Moon surface	8.5623	24.0696	15.3094	25.0325	**25.5287**
Stream and bridge	14.5846	21.9104	16.7949	**27.7772**	21.0173
Tank	7.3563	16.9021	11.5328	19.5906	**28.0855**
Tank2	9.6327	27.8177	18.1442	**37.5682**	32.0589
Tank3	9.1506	17.7282	12.5414	16.6483	**20.3116**
Truck and APCs	10.6206	24.9340	17.0698	**30.2223**	23.6833
Truck and APCs2	10.5282	22.8365	15.8909	**32.4396**	24.2075
Truck	6.8949	15.4893	10.8829	19.9877	**20.4818**

TABLE V. LINEAR INDEX OF FUZZINESS (IOF) AND ENTROPY TEST RESULTS

	IOF		Entropy	
Image	Original	Proposed	Original	Proposed
APC	0.7640	**0.6601**	5.0534	**5.5069**
Aerial	0.6868	**0.4045**	**7.3118**	6.8089
Aerial2	0.5362	**0.2484**	**6.9940**	6.0122
Airplane (U-2)	0.2751	**0.1507**	**5.6415**	4.9032
Airplane	0.3137	**0.2481**	**6.4523**	5.0018
Airplane2	0.5756	**0.2566**	4.0045	**4.3660**
Airport	0.6210	**0.4144**	**6.8303**	6.7218
Car and APCs	0.7875	**0.6152**	6.1074	**6.8027**
Car and APCs2	0.7743	**0.6124**	5.9088	**6.8092**
Chemical plant	0.6534	**0.4826**	7.3424	**7.4994**
Clock	0.3552	**0.1479**	**6.7057**	4.1180
Couple	0.7660	**0.5440**	7.2010	**7.6298**
Fishing Boat	0.7130	**0.4747**	7.1914	**7.2796**
Male	0.5614	**0.4259**	**7.5237**	7.5131
Moon surface	0.8306	**0.5784**	6.7093	**7.4148**
Stream and bridge	0.6279	**0.3841**	5.7056	**6.9006**
Tank	0.7407	**0.5702**	5.4957	**6.2501**
Tank2	0.8360	**0.6383**	5.9916	**6.9583**
Tank3	0.7367	**0.4749**	6.1898	**6.8043**
Truck and APCs	0.7151	**0.5405**	6.5632	**7.2792**
Truck and APCs2	0.6739	**0.5547**	6.6953	**7.2057**
Truck	0.8096	**0.6486**	6.0274	**6.6620**

'Aerial', 'Airplane', and 'Stream and bridge', the proposed method achieves a lower MSE value, and a higher PSNR and SSIM value. Table IV shows that in more than half of the test images, the proposed method obtained a higher PSNR value. In Table V, compared to the original test images, all the result images have a lower IOF value, and More than half of the result images achieve a higher entropy value. The above experimental results show that the proposed method has a good enhancement effect.

IV. Conclusion

This paper proposed an image enhancement method base on an improved fuzzy c-means clustering(FCM-co). The FCM-co has a cooperation center and it could consider image pixels' location information. The paper introduces a new function to modify the membership value. Through comparative experiments, the results show that the proposed method has a good enhancement effect. In the following work, we intend to try to change the value of parameters c and α for further research. We also plan to apply FCM-co to other areas, such as image segmentation.

Acknowledgment

The authors would like to express their appreciation and gratitude to the Research Management Centre, Universiti Malaysia Sabah for granting this research study under Skim UMSGreat (GUG0540-2/2020).

References

[1] R. C. Gonzalez, R. E. Woods, and B.R. Masters, *Digital Image Processing*, 3rd ed, 2009.

[2] R. Hummel, *Image enhancement by histogram transformation*, Computer Graphics and Image Processing. 1977, vol. 6, no. 2, p. 184-195.

[3] S. C. F. Lin, C. Y. Wong, M. A. Rahman, G. Jiang, S. Liu, N. Kwok, H. Shi, Y. H. Yu, T. Wu, *Image enhancement using the 160 averaging histogram equalization (AVHEQ) approach for contrast improvement and brightness preservation*, Computers and 161 Electrical Engineering. 2015, vol. 6, no. 356-370.

[4] G. Ulutas and B. Ustubioglu, *Underwater image enhancement using contrast limited adaptive histogram equalization and layered difference representation*, Multimedia Tools and Applications. 2021, vol. 80, no. 2, 10, p. 15067-15091.

[5] S. Kumar, A. K. Bhandari, A. Raj and K. Swaraj, *Triple clipped histogram-based medical image enhancement using spatial frequency*, IEEE Transactions on NanoBioscience, 2021, vol. 20, no. 3, p. 278-286.

[6] S. C. Huang, F. C. Cheng, and Y. S. Chiu, *Efficient contrast enhancement using adaptive gamma correction with weighting distribution*,IEEE transactions on image processing. 2012, vol. 22, no. 3, p. 1032-1041.

[7] A. Kumar, R. K. Jha and N. K. Nishchal, *An improved Gamma correction model for image dehazing in a multi-exposure fusion framework*, Journal of Visual Communication and Image Representation. 2021, vol. 78, p.103-122.

[8] S. Liu, W. Long, L. He, Y. Li and W. Ding, *Retinex-based fast algorithm for low-light image enhancement*, Entropy. 2021, vol. 23, no. 6, p. 746.

[9] L. S. S. Singh, A. K. Ahlawat, D. K. M. Singh, and T. P Singh, *Image Enhancement Techniques On Fuzzy Domain: A Review*, International Journal of Computer Engineering and Technology. 2017, vol. 8, no. 2, p. 80-99.

[10] H. Deng, C. Duan, X. Zhou, *A novel fuzzy enhancement of mammograms*, 2015 IET International Conference 177 on Biomedical Image and Signal Processing (ICBISP 2015), November 2015, pp. 1-5.

[11] L. Yang, S. Zenian, R. Zakaria, *Fuzzy image enhancement based on algebraic function and cycloid arc length*, 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), September 2021, p. 1-4.

[12] H. D. Cheng and H. Xu, *A novel fuzzy logic approach to contrast enhancement*, Pattern recognition. 2000, vol. 33, no. 5, p. 809-19.

[13] M. Hanmandlu, O. P. Verma, N. K. Kumar and M. Kulkarni, *A novel optimal fuzzy system for color image enhancement using bacterial foraging*, IEEE Transactions on Instrumentation and Measurement. 2009, vol. 58, no. 8, p. 2867-2879.

[14] G. Li, Y. Tong and X. Xiao, *Adaptive Fuzzy Enhancement Algorithm of Surface Image based on Local Discrimination via Grey Entropy*, Elsevier Procedia Engineering. 2011, vol. 15, p. 1590 C 1594.

[15] K. Hasikin and N. A. M. Isa, *Enhancement of the low contrast image using fuzzy set theory*, 14th International Conference on Modelling and Simulation, March 2012, p. 371-376.

[16] A. K. Gupta, S. S. Chauhan and M. Shrivastava, *Low contrast image enhancement technique by using fuzzy method*, International Journal of Engineering Research and General Science. 2016, vol. 4, no. 2, p. 518-526.

[17] V. Magudeeswaran and J. F. Singh, *Contrast limited fuzzy adaptive histogram equalization for enhancement of brain images*, International Journal of Imaging Systems and Technology. 2017, vol. 27, no. 1, p. 98-103.

[18] S. Zenian, T. Ahmad, and A. Idris, *Intuitionistic fuzzy set: FEEG image representation*, AIP Conference Proceedings 2184,060054(2019), 2019.

[19] S. Zenian, T. Ahmad, and A. Idris, *Fuzzy contrast enhancement by intensification operator in Flat Electroencephalography Image*, Transactions on Science and Technology. 2019, vol. 6, no. 2-2, p. 216-220.

[20] J. C. Dunn, *A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters*, Journal of Cybernetics. 1973, vol. 3, p. 32-57.

[21] J. C. Bezdek, *Objective function clustering*, Pattern recognition with fuzzy objective function algorithms, Springer, 1981, p. 43-93.

[22] D. Naik and P. Shah, *A review on image segmentation clustering algorithms*, Int J Comput Sci Inform Technol. 2014, vol. 5, no. 3, p. 3289-93.

[23] L. Chen, Z. Li, Z. Li, S. Chen, Q. Yang and Y. Du, *A Contrast Enhancement Method of Infrared Finger Vein Image Based on Fuzzy Technique*, 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), November 2019, p. 307-310.

[24] T. Chaira and A. K. Ray, *Fuzzy image processing and applications with MATLAB*, CRC Press, 2017.

[25] R. Kumar and A. K. Bhandari, *Fuzzified Contrast Enhancement for Nearly Invisible Images*, IEEE Transactions on Circuits and Systems for Video Technology. 2021, vol. 32, no. 5, p. 2802-2813.

[26] M. Hanmandlu, D. Jha and R. Sharma, *Color image enhancement by fuzzy intensification*, Pattern recognition letters. 2003, vol. 24 no. 1-3, p. 81-87.

[27] R. C. Gonzalez, R.E. Woods and S.L. Eddins, *Digital Image Processing Using MATLAB*, New Jersey, Prentice Hall, 2003.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, *Image quality assessment: from error visibility to structural similarity*, IEEE transactions on image processing. 2004, vol. 13 no. 4, p. 600-612.

A New Hate Speech Detection System based on Textual and Psychological Features

Fatimah Alkomah, Sanaz Salati, Xiaogang Ma
Department of Computer Science
University of Idaho
Moscow, ID

***Abstract*—Hate speech often spreads on social media and harms individuals and the community. Machine learning models have been proposed to detect hate speech in social media; however, several issues presently limit the performance of current approaches. One challenge is the issue of having diverse comprehensions of hate speech constructs which will lead to many speech categories and different interpretations. In addition, certain language-specific features, and short text issues, such as Twitter, exacerbate the problem. Moreover, current machine learning approaches lack universality due to small datasets and the adoption of a few features of hateful speech. This paper develops and builds new feature sets based on frequencies of textual tokens and psychological characteristics. Then, the study evaluates several machine learning methods over a large dataset. Results showed that the Random Forest and BERT methods are the most valuable for detecting hate speech content. Furthermore, the most dominant features that are helpful for hate speech detection methods combine psychological features and Term-Frequency Inverse Document-Frequency (TFIDF) features. Therefore, the proposed approach could identify hate speech on social media platforms like Twitter.**

Keywords*—Hate speech detection; hate speech classification; hate speech features; hate speech methods

I. INTRODUCTION

As the number of users of social media increases, the impact of hate speech is drastic due to the ease of posting hate speech without geographical boundaries and user anonymity. The uncontrolled spread of hate can damage our society gravely and severely harm marginalized people or groups [1]. The effect of hate crimes is widely spread due to the users' anonymity[2] and the wide use of social media. Twitter, as social media, was studied by 54.81% of researchers; primarily, textual analysis was the prevalent method with 33% compared to other methods [3].

Hate speech detection is a challenging research problem due to many issues, including competing definitions, limited feature sets, small-sized datasets, and the current design of current models. Competing hate speech definitions capture different information with different interpretations by proposed models. For example, racist and homophobic tweets are more likely to be classified as hate speech. However, some definitions are debatable [4]. Therefore, the nonexistence of a universally accepted definition is due to whether offensive conveys hate or not [5]. The critical aspect is separating hate speech language from other offensive languages [6]. The problem of competing definitions would result in a poor feature detection set that could not help identifying hate speech. The problem posed by ungrammatical text has mainly been used to mitigate the difficulty of automatically detecting hateful speech, particularly when users intentionally change keywords' spelling or avoid automatic content [7], [8].

The issue of feature detection becomes more challenging as some words are contextual dependent on users and groups and are not inherently offensive [9], [10]. Small-sized datasets are not enough to generalize results or capture compelling hate speech detection features. For example, Cervero's method [11] employs 200 tweets and yet achieves a good result. Obstacles also include partially labeled data, which makes comparing the performance of many datasets hard to validate. Therefore, many machine learning models do not generalize any hate speech content as it is limited to specific keywords or dictionaries [11]. For example, it was shown that the Yin and Zubiaga model's performance[12] drops down by 10% when tested on another dataset outside the same group of datasets. As a result, the feature sets of datasets do not necessarily represent real-life cases, despite reported performance[11]. Therefore, several machine learning models cannot scale well in practice or models that are not robust due to dataset bias.

This paper develops several machine learning models that are helpful in detecting hate speech based on textual tweets on Twitter. The paper uses the Twitter dataset of 150k tweets [13], called the MMHS150K benchmark dataset. The images were removed from the dataset, and the dataset was converted from the JSON to a tabular format. Three textual features were extracted from the dataset: the frequency of user mentions, hashtags, and emojis; TFIDF of 3-grams; and psychological features extracted by the Linguistic Inquiry and Word Count (LIWC) [14]. Linguistic Inquiry and Word Count is a software application for counting words that references a lexicon of grammatical, psychological, and content word categories. LIWC has been used to categorize texts effectively along psychological dimensions (such as users' personality traits and emotions). The proposed approach was tested on Naïve Bayes, Gradient Boosting, XGBoost, Random Forest, KNN, and Decision Trees algorithms. This study aims to present a model that could be used to automate hate speech detection on any social media platform such as Twitter. We also aim to find the best features that work well with the best-performing algorithm.

The proposed method has several contributions aside from using existing machine learning models from conventional and deep learning methods. This study has extensively studied

the effect of three different groups of feature sets on the results of hate speech detection. We have shown that combining more than one feature set provides a good performance model. Moreover, the proposed method studies the multilabel classification problem and delivers results at the label level, which was lacking in previous studies. Additionally, the proposed model could be integrated with social media platforms to instantly detect and block hate speech.

Our research objectives include identifying textual features that were effective in the classification. For example, the model should be able to detect hate speech fine-grained at the label level given a short text (Tweet).

The paper is outlined as follows. Related works are summarized in Section II. Section III illustrates the proposed machine learning approach. Results and discussions are explained in Section IV and V. The paper is concluded in Section VI.

II. Related Work

Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) is a new track to detect hate speech detection in the research community. The HASOC track intends to provide a platform to develop and optimize Hate Speech detection algorithms for Hindi, German and English [15]. The best result on the English language dataset of HASOC was based on Long Short-Term Memory (LSTM), which used GloVe embeddings as input. The best system achieved a performance of f1-measure of 0.52; however, the dataset has only 3,708 records for the English dataset. The International Workshop on Semantic Evaluation (SemEval) organizes the OffensEval series of shared tasks on offensive language identification using the hierarchical annotation of the type and target of offensive content [16] [17]. However, robust datasets survive in many classification tasks of hate speech and are reusable and easy to update.

Furthermore, it was reported that robust datasets are required to allow comparability of features and methods [18]. Therefore, as posts of hate speech can also be implicit, few lexical features could be used for machine learning models. Although there are many approaches and features, the current list of models cannot be generalized due to dataset size, credibility, low precision, or imbalanced datasets.

The literature reported various features of hate speech that include shallow lexical features [19], dictionaries [20], sentiment analysis [21], linguistic characteristics [22], knowledge-based features [23], and meta-information [24] of social media content. Readers may refer to a comprehensive study of hate speech detection methods and datasets published recently [25]. However, the literature showed that shallow lexical detection methods have low precision [19]. The literature reported that identifying hate speech on a large scale is still an unsolved problem [26]. For example, the DeepHate method [16] is based on many features: word embeddings, sentiment, and topic information. Recently, aggressive and gendered identification are getting attention [27]. It was found that stylometric (such as function words) and emotion-based features are robust indicators of hate speech[28]. Markov *et al.* [28] provided a model based on encoded emotion information of 14,182 emotion words and their association with emotions and sentiments from the emotion lexicon [29]. Furthermore, the Linguistic Inquiry and Word Count (LIWC) of Pennebaker *et al.* [14] and profanity [30] (especially anger) are good indicators of hate speech in the Indian language context [31]. The LIWC categories include linguistic statistics such as counts and summary variables: analytic, clout, authenticity, and emotional tone. In addition, the LIWC could reveal feelings, personality, and psychological motivations [14]. However, it was shown that the features relating to users' personality traits and emotions in text achieved an accuracy result of 0.7 in English text [32].

Therefore, current methods lack a suitable set of features for hate speech; are either based on small datasets or have low performance when tested over multiclassification hate speech problems. The overall issue is related to the nonexistence of a universally accepted definition of hate speech which results in whether offensive tweets convey hate or not [5].

III. Proposed Framework

In this study, the proposed framework is a machine learning model with an input of a hate speech dataset and trained binary classification output. The framework (Fig. 1) has four steps: data preparation, feature extraction, model learning, and classification output.

A. Data Preparation

It was found that datasets target multiple hate speech categories; however, only 60% of dataset builders reported an inter-annotator agreement [33]. Moreover, it is common for many datasets to overlap between class labels, as Waseem [34] showed an overlap of 2,876 tweets with the Waseem and Hovy datasets [35]. Therefore, relevant and no obsolete datasets are essential to a useful predictive hate speech model. However, creating large and varied hate or abusive datasets that minimize potential bias is laborious and requires specialized experts [36]. Therefore, this study uses a large benchmark dataset taken from a previous Twitter dataset of 150k tweets [13], the MMHS150K dataset. The dataset has an average tweet length of 91 characters, a minimal length of 15, and a maximum length of 193, including the URLs. The dataset has images and textual data of tweets and image captions from Twitter in a python dictionary inside a JSON file. The key of each entry in the JSON file is the tweet ID. The other fields include three different fields, which are the image URL, tweet URL, tweet text, and class labels. The dataset has six classes, shown in Table. I.

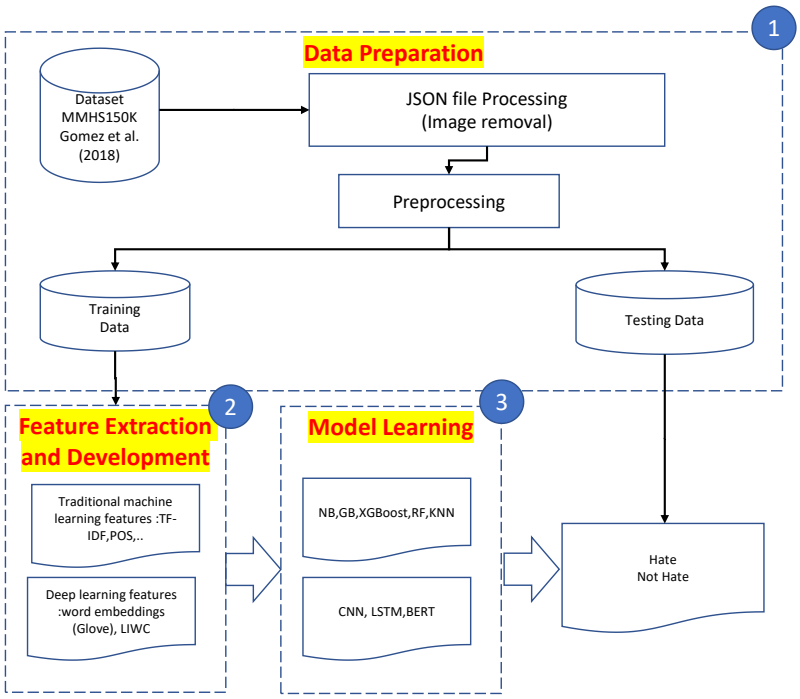


Fig. 1. Proposed Framework.

TABLE I. DISTRIBUTION OF CLASS LABELS IN THE BENCHMARK DATASET*

Class	Total Instances
Not Hate	131,081
Racist	44,535
Sexist	19,509
Homophobe	10,554
Religion	2,119
Other Hate	21,217
Total	143,277

* Further details about extracting and preparing the original dataset are found here https://gombru.github.io/2019/10/09/MMHS/

B. Data Preprocessing

The following are the text preprocessing actions carried out in this study.

1) Removal of images and keeping only textual content in the dataset. This step involves converting the dataset into a tabular format for further preprocessing.

2) Stop words removal.

3) Convert text to lowercase after counting the number of capital letter words.

4) Removal of user mention after checking if a tweet has a mentioned user.

5) Emotions extraction using the UNICODE_EMOJI library from the emot.emo_unicode package.

6) Convert emojis to placeholders so that they will be part of the 3-grams.

7) Tokenization.

8) Lemmatization.

9) 3-grams Extraction.

*10)*Convert text to TF-IDF vector.

C. Feature Extraction and Development

Based on previous literature, this study selects several feature sets such as frequency of tokens (e.g., hashtags) or TFIDF and word embeddings. We follow the following criteria for selecting the sets of features: (1) features must be used in prior hate speech detection models with evidence of acceptable results, (2) the feature must be textual and in line with the current dataset characteristics, and (3) the feature should be used by at least two related studies. Therefore, following these criteria, the features are explained in Table II.

Notably, the selection of feature set 3 is used by only one related study; however, such feature set (LIWC) was evident in other studies related to human sentiments. Therefore, different combinations of the three groups will be used with various machine learning algorithms.

D. Model Learning

This study examines the performance of traditional and deep learning methods on the benchmark dataset. A good model must use the minimum number of features; therefore, this study finds the best features that maximize performance. Consequently, the following methods were selected from machine learning: Naïve Bayes, Gradient Boosting, XGBoost, Random Forest, KNN, and Decision Trees. The benchmark dataset was split into training and testing (80% training and 20% for testing). Stratified sampling is used to ensure proper sampling for each class label. The dataset is imbalanced; therefore, the dataset is balanced using oversampling techniques of SMOTE, where BorderlineSMOTE was the best.

TABLE II. THE FEATURE SET FOR HATE SPEECH DETECTION

Category	Name	Description	Rationale	Related Studies
Feature Set 1 (Counts)	Username Mention	Checks if a tweet mentioned any other user. The preprocessing uses '1' if a username is mentioned, '0' otherwise	Mentioning a person could indicate hate toward that person.	[37]
	Capital Letter	Count of words with Capitals letters.	Tweets that have capital letters may indicate hate or stress in speech.	[38]
	Hashtags	Count of hashtags	Similar to the user mention	[37] [39] [40]
	Emojis	Count of emojis in the tweets	Could indicate a negative attitude of users	[26]
Feature Set 2 (TFIDF)	TFIDF Or Word Embeddings	Cleaned TFIDF vector. It includes tokenization, stemming, and 3-grams. Word embeddings (Glove) is used with Deep learning models.	The semantic structure of tweets	[41] [42] [37] [43] [44] [45]
Feature Set 3 (LIWC)	Summary Language Variables	analytic, clout, authentic, and tone	Analytical thinking, clout, authenticity, emotional tone	Proposed by this study
	Linguistic Dimensions	i, we, you, shehe, they, and ipron	It was shown that the hate speech features relating to users' personality traits and emotions in text achieved an accuracy result of 0.7 in English text [32].	Proposed by this study
	Psychological Processes	Affect variables posemo, negemo, anx, anger, and sad		[46]
	social	family, friend, female, and male		Proposed by this study
	bio	body, health, sexual, and ingest		Proposed by this study
	drives	affiliation, achieve, power, reward, and risk		Proposed by this study

The results that predict hate speech are anlayzed following the standards of machine learning. The precision, recall, f1-measure, and ROC standard performance metrics. The analysis also includes a comparison with other methods in hate speech detection. The analysis also includes identifying the most predictable features for each model.

IV. EXPERIMENTS, RESULTS AND DISCUSSION

Following the selected features in Table II and after preprocessing the dataset, each feature set was created using the python scikit-learn library. Each feature set was prepared alone, allowing different feature sets to be combined with various machine learning algorithms. For example, the first feature sets were processed as follows:

1) If a Tweet includes a usernames, add another feature that is '0' has does not include a username and and '1' has mentioned any user name.

2) If the Tweet has capital letter words, add a new feature and place the total number of capitalized words.

3) Keep hashtags as they might be informative.

4) Replacing URLs, mentions, emojis, Retweets, and capital letters as placeholders.

For example, the tweet

'@Mr Rodie94 Nigga was in the store like 😂 https://t.co/dSEb83kIhm'.

becomes

'mention placeholder nigga was in the store like face_with_tears_of_joy url_placeholder'.

The second feature set is the TFIDF with 3-grams, which was carried out using python; the top trigrams are shown in Fig. 2. Samples of preprocessing steps are shown in Fig. 3. Finally, the third feature set was extracted with LIWC software, which in turn was exported to excel and preprocessed with python.

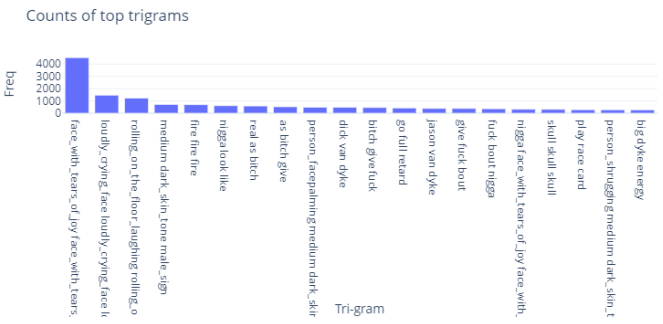


Fig. 2. Top Trigrams.

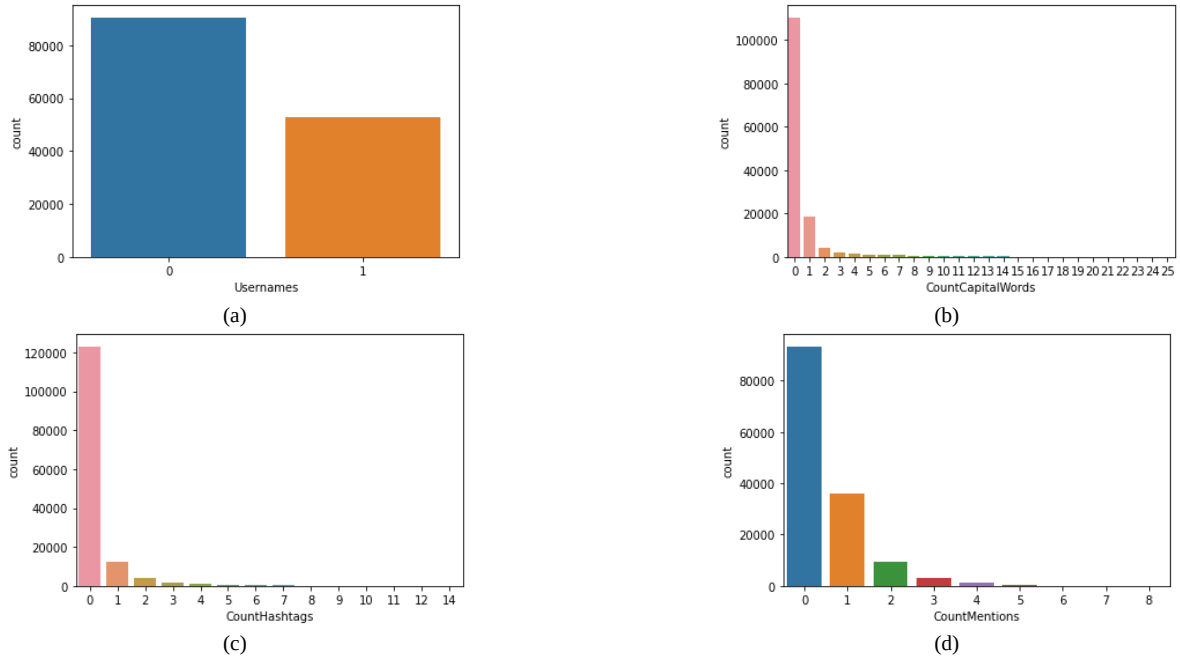


Fig. 3. Preprocessing Steps. Showing Usernames for Both Negative and Positive Example (a), Counting of Capital Words (b), Counting of Hashtags(c), and Count of Mentioned (d).

E. Application of the Proposed Methods (Model Learning)

The classification of this research is a binary classification where each machine learning algorithm is tested on the dataset (hate/not hate). The adopted methods are explained in Table III. On the other hand, the deep learning structure for binary classification of hate speech is shown in Appendix A. The parameters were deduced as per many experiments considering that the nature of machine learning is multiclassification. Each feature set was first to run alone with a specific method, and then the features were combined together.

TABLE III. TRADITIONAL MACHINE LEARNING METHODS FOR BINARY CLASSIFICATION OF HATE SPEECH

Algorithm	**Settings**
Naïve Bayes	naive_classifier = MultinomialNB()
Gradient Boosting	criterion='friedman_mse', init=None, learning_rate=0.1, loss='deviance', max_depth=3, max_features='log2', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=80, n_iter_no_change=None, random_state=None, subsample=1.0, tol=0.0001,
	validation_fraction=0.1
XGBoost	base_score=0.5, booster='gbtree', colsample_bytree=0.6, gamma=0.3, learning_rate=0.01, max_depth=3, min_child_weight=1, n_estimators=20, random_state=40, reg_alpha=0, reg_lambda=1.5, scale_pos_weight=1, seed=None, subsample=0.4
Random Forest	n_estimators = 200
KNN	algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=None, n_neighbors=10, p=2, weights='uniform'
Decision Trees	riterion='entropy', random_state=1
Deep Learning models	
CNN	Structure varies based on feature sets, as explained in A ppendix A
LSTM	Structure varies based on feature sets, as explained in A ppendix A
BERT	Structure varies based on feature sets, as explained in A ppendix A

F. Classification Output Analysis

Following the machine learning Table III, Appendix A, and the proposed set of features in Table II, the results are depicted in Fig. 4-6 and discussed here. As shown in Fig. 4, the first feature set is the lowest-performing feature set, indicating that such features are not performing well. However, the second and the third feature sets provide promising results with the most studied algorithms. The highest performance was for the BERT, with a 0.974 f1-score measure on the second feature set and 0.956 on both the first and the third feature sets. The f1-measure for positive and negative examples of the selected machine learning models is shown in Fig. 5. The figure shows that models provide high performance for positive examples (hate=1) and low performance for negative examples. This finding is consistent with previous works [19] and shows that negative examples are still challenging due to the use of similar keywords, as illustrated earlier [6] [5]. Therefore, the nonexistence of a universally accepted definition is due to whether offensive conveys hate or not [5]. Overall, the proposed model provided higher performance in binary classification, 0.98 compared to the original model of a maximum of 0.734 [47].

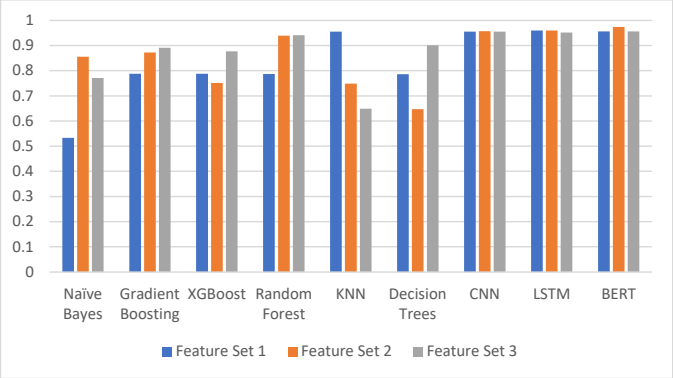


Fig. 4. Traditional and Deep Machine Learning Algorithms F1-Measure against Feature Sets.

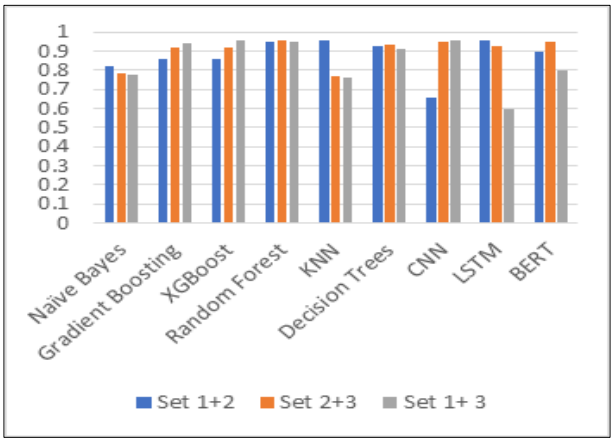


Fig. 5. Selected Algorithms Average F1-Measure (Binary Classification Feature Combinations).

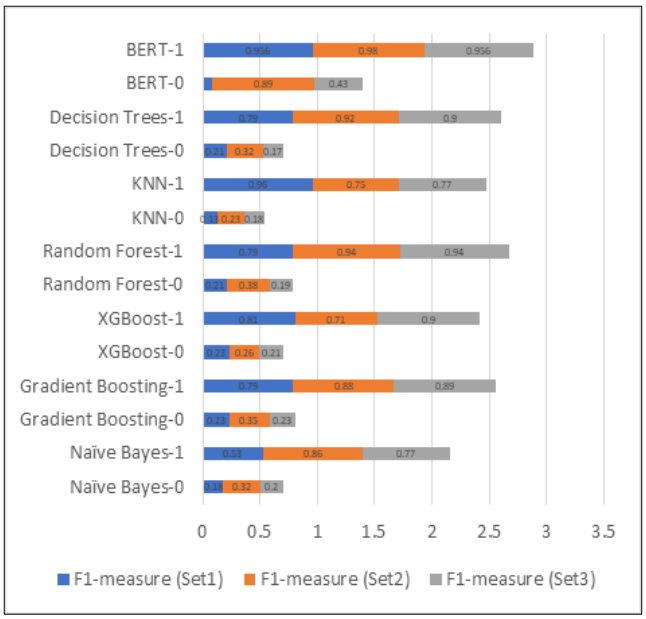


Fig. 6. Deep Learning Machine Learning Algorithms F1-Measure against Feature Sets.

Next, the performance of LSTM, CNN, and BERT (along with the baseline methods) are shown in Figure 6. For BERT: bert_multi_cased_L-12_H-768_A-12/2 model were used. The f1-measure for the BERT model is the highest among the deep learning models. The structure of these algorithms is shown in Appendix A. As compared with previous methods, BERT is the most promising method. The reported f1-measure for BERT is 0.974. Above all, BERT was the most prominent method that distinguishes the negative examples of hate speech, as shown in Fig. 5. However, in practice, it is essential to select the best performing set of features that provides the optimal model. Fig. 5 shows the list of selected algorithms and their performance when several features are merged. It shows that the LSTM got an f1-measure of 0.96 when combining features (set 1 and set 2). Contrary to our previous finding that BERT is the best, it was not performing as compared to LSTM due to the complexity of integrating feature sets of the original bert_multi_cased model and the new features extracted from text. Nevertheless, the most consistent algorithm for the random forest provides relatively similar results when different feature sets.

Table IV shows a sample of related works on hate speech classification and their performance. Unfortunately, most models are not available to the public and were tested in different datasets. Therefore, careful interpretation of the results in the table should be considered as different datasets will eventually change the model outcomes; this issue is already discussed before.

TABLE IV. TRADITIONAL MACHINE LEARNING METHODS FOR BINARY CLASSIFICATION OF HATE SPEECH

Ref	Dataset	Best Method	Accuracy	F1-measure
[48]	Islamophobic hate speech:100K tweets	One-versus-one SVM	0.77	
[19]	25K tweets	SVM		**0.91**
[49]	5K tweets	Logistic Regression	0.704	
[39]	76 K tweets	MCD + LSTM	0.78	
[50]	6.6K tweets	GRU + CNN		**0.78**
[51]	14 K for SemEval-2019 Task 6 subtask A: Offensive/non-offensive	MCD + LSTM	0.78 F1-score	
[52]	SemEval-2019 Task 6 [154]	GRU + CNN	Task A: classification of tweets into either offensive (OFF) or not offensive (NOT) 0.78 for supervised 0.77 for the unsupervised approach	
[53]	**six datasets and 121 customized list**	**Cat Boost**		**F1-score ranging from 0.85 to 0.89 Best average F1-score 87.74 across all datasets**

V. DISCUSSION

Our research objectives include identifying textual features that were effective in the classification. The research showed that the most dominant features are textual features extracted from TFIDF features, as shown in Fig. 2. The features are focused on emotional features such as face_with_tears_of_joy, which was evident in the dataset with 4,528 frequent items. In addition, other keywords were frequent, such as 'fire,' 'nigaal', 'dick van dyke', and others. Such a finding is consistent with previous studies that showed that sentiments are effective in showing a large number of hate speech contents [37], [38], [54]. In addition, the findings are consistent with works related to LIWC as additional features showing human behavior [46].

The developed machine learning models showed that, as expected, the binary classification was providing acceptable results. The best performing model was BERT with 0.974. LSTM also reported good results with an f1-measure of 0.96. The reason is that these models depend on high-dimensional word embedding, and their design was proved to work well with many textual classification tasks. The combination of feature set 2 and feature set 3 provides good results for LSTM and BERT models. The other models reported lower performance, such as CNN (below 0.66 f1-meaure) for the combination of feature set 1 and feature set 2. A single feature set, such as feature set 2 performed well on most algorithms. The best-performing model reported an f1-score of binary classification f1-score of 0.704 with the Feature Concatenation Model (FCM) [13]. The proposed model reported LSTM with an f1-measure of 0.96 (feature set1+feature set 2) with binary classification and 0.96 on LSM and CNN (feature set 2+feature set 3). However, the proposed model has not reported good performance for each label. The investigations showed the original imbalanced dataset, which does not have enough examples for each label. Due to the complexity of hate speech detection, decision trees and KNN provided high f1-measure performance based on TFIDF feature sets. However, these algorithms did not generalize well at the label level (hate/not hate), indicating that there were standard features between positive and negative examples of the hate speech benchmark dataset.

Consequently, with a wide set of machine learning models, the results indicate that as the number and type of features are added (shown in groups in Table II.), the machine learning model performance increase. The reason is that the additional features add new semantics to the embedded or intended meaning in a particular Tweet. For example, the LIWC features (Feature set 3), have shown relatively good performance in detecting sentiments and user psychological features.

Although the experiments have been run on a single dataset, the dataset is considered one of the largest datasets that are available online. According to a previous study, it was found that current datasets suffer from various aspects, including their size, bias, and authenticity in terms of the annotation process [25]. A comparison of hate speech models was not fully available as many models are not published, or the dataset is private. However, the proposed model was able to provide an acceptable accuracy with a baseline work that used additional non-textual features such as images and their captions [13]. Therefore, given these restrictions, and due to the complexity of hate speech features, the results are considered acceptable but should be interpreted within the context of hate speech categories implied in the adopted benchmark dataset. The new work provides implications to theory with newly adapted machine learning models and could be used on unseen data on Twitter or similar social media platforms.

VI. CONCLUSION

This paper develops three feature sets that could be used for hate speech detection: frequencies of unique tokens, TFIDF, and LIWC features. Then, the paper extensively compares several machine learning models: Naïve Bayes, Gradient Boosting, XGBoost, Random Forest, KNN, Decision Trees, LSTM, CNN, and BERT. The difficulty of hate speech identification was shown by the high f1-measure performance of decision trees and KNN based on TFIDF feature sets. However, these algorithms did not generalize effectively at the label level (hate/not hate), showing that positive and negative samples of the hate speech benchmark dataset shared common

characteristics. Conversely, the results of the BERT model were relatively higher, with an f1-measure of 0.974 on the same feature set (TFIDF). In addition, the LIWC feature sets and their combination with TFIDF provided better results on the LSTM method. However, features among the adopted LIWC could share common information. It is recommended that the adopted approach should be considered in the context of generic hate speech on a short text like Twitter. The model might need retraining due to out-of-vocabulary keywords that users might use over time. Furthermore, the researchers might consider another resource of hate speech aside from Twitter. Therefore, we plan to test the models based on a single sub feature on a leave-out scheme in the future.

REFERENCES

[1] M. Bedrosova, H. Machackova, J. Šerek, D. Smahel, and C. Blaya, "The relation between the cyberhate and cyberbullying experiences of adolescents in the Czech Republic, Poland, and Slovakia," Comput. Human Behav., vol. 126, p. 107013, 2022, doi: https://doi.org/10.1016/j.chb.2021.107013.

[2] S. T. Peddinti, K. W. Ross, and J. Cappos, "User anonymity on twitter," IEEE Secur. Priv., vol. 15, no. 3, pp. 84–87, 2017.

[3] A. Matamoros-Fernández and J. Farkas, "Racism, Hate Speech, and Social Media: A Systematic Review and Critique," Telev. New Media, vol. 22, no. 2, pp. 205–224, 2021, doi: 10.1177/1527476420982230.

[4] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine learning techniques for hate speech classification of twitter data: State-of-The-Art, future challenges and research directions," Comput. Sci. Rev., vol. 38, p. 100311, 2020, doi: 10.1016/j.cosrev.2020.100311.

[5] N. Strossen, "Freedom of speech and equality: Do we have to choose," JL Pol'y, vol. 25, p. 185, 2016.

[6] S. MacAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," PLoS One, vol. 14, no. 8, pp. 1–16, 2019, doi: 10.1371/journal.pone.0221152.

[7] A. Nourbakhsh, F. Vermeer, G. Wiltvank, and R. van der Goot, "sthruggle at SemEval-2019 Task 5: An Ensemble Approach to Hate Speech Detection," pp. 484–488, 2019, doi: 10.18653/v1/s19-2086.

[8] G. Kovács, P. Alonso, and R. Saini, "Challenges of Hate Speech Detection in Social Media," SN Comput. Sci., vol. 2, no. 2, pp. 1–15, 2021, doi: 10.1007/s42979-021-00457-3.

[9] S. Ullmann and M. Tomalin, "Quarantining online hate speech: technical and ethical perspectives," Ethics Inf. Technol., vol. 22, no. 1, pp. 69–80, 2020, doi: 10.1007/s10676-019-09516-z.

[10] E. Mosca, M. Wich, and G. Groh, "Understanding and Interpreting the Impact of User Context in Hate Speech Detection," no. Ml, pp. 91–102, 2021, doi: 10.18653/v1/2021.socialnlp-1.8.

[11] S. D. Swamy, A. Jamatia, and B. Gambäck, "Studying generalisability across abusive language detection datasets," in Proceedings of the 23rd conference on computational natural language learning (CoNLL), 2019, pp. 940–950.

[12] W. Yin and A. Zubiaga, "Towards generalisable hate speech detection: a review on obstacles and solutions," PeerJ Comput. Sci., vol. 7, pp. 1–38, 2021, doi: 10.7717/PEERJ-CS.598.

[13] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," Proc. - 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV 2020, pp. 1459–1467, 2020, doi: 10.1109/WACV45572.2020.9093414.

[14] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," 2015.

[15] T. Mandl et al., "Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in Indo-European languages," CEUR Workshop Proc., vol. 2826, pp. 87–111, 2020.

[16] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 1415–1420, 2019, doi: 10.18653/v1/n19-1144.

[17] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, "SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification," Find. Assoc. Comput. Linguist. ACL-IJCNLP 2021, pp. 915–928, 2021, doi: 10.18653/v1/2021.findings-acl.80.

[18] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," no. 2012, pp. 1–10, 2017, doi: 10.18653/v1/w17-1101.

[19] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," Proc. 11th Int. Conf. Web Soc. Media, ICWSM 2017, no. Icwsm, pp. 512–515, 2017.

[20] V. Lingiardi, N. Carone, G. Semeraro, C. Musto, M. D'Amico, and S. Brena, "Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis," Behav. Inf. Technol., vol. 39, no. 7, pp. 711–721, 2020, doi: 10.1080/0144929X.2019.1607903.

[21] F. H. A. Shibly, U. Sharma, and H. M. M. Naleer, Classifying and Measuring Hate Speech in Twitter Using Topic Classifier of Sentiment Analysis, vol. 1165. Springer Singapore, 2021. doi: 10.1007/978-981-15-5113-0_54.

[22] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," 12th Int. AAAI Conf. Web Soc. Media, ICWSM 2018, no. ICWSM, pp. 42–51, 2018.

[23] H. Abburi, S. Sehgal, and H. Maheshwari, "Knowledge-based Neural Framework for Sexism Detection and Classification," no. September, 2021.

[24] F. Rangel, G. L. D. L. P. Sarracén, Bert. Chulvi, E. Fersini, and P. Rosso, "Profiling Hate Speech Spreaders on Twitter Task at PAN 2021," CLEF 2021 Labs Work. Noteb. Pap., no. September, pp. 21–24, 2021.

[25] F. Alkomah and X. Ma, "A Literature Review of Textual Hate Speech Detection Methods and Datasets," Information, vol. 13, no. 6, 2022, doi: 10.3390/info13060273.

[26] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "A Multilingual Evaluation for Online Hate Speech Detection," ACM Trans. Internet Technol., vol. 20, no. 2, 2020, doi: 10.1145/3377323.

[27] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Evaluating Aggression Identification in Social Media," Proc. Second Work. Trolling, Aggress. Cyberbullying, no. May, pp. 1–5, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.trac-1.1.

[28] I. Markov, N. Ljubesic, D. Fiser, and Walter, "Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection," Proc. 11th Work. Comput. Approaches to Subj. Sentim. Soc. Media Anal., pp. 149–159, 2021, [Online]. Available: https://www.aclweb.org/anthology/2021.wassa-1.16/.

[29] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word--emotion association lexicon," Comput. Intell., vol. 29, no. 3, pp. 436–465, 2013.

[30] T. Jay and K. Janschewitz, "The pragmatics of swearing," 2008.

[31] P. Mathur, R. Sawhney, M. Ayyar, and R. Shah, "Did you offend me? Classification of Offensive Tweets in Hinglish Language," pp. 138–148, 2019, doi: 10.18653/v1/w18-5118.

[32] R. Cervero, "Use of Lexical and Psycho-Emotional Information to Detect Hate Speech Spreaders on Twitter," 2021.

[33] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: a systematic review," Lang. Resour. Eval., vol. 55, no. 2, pp. 477–523, 2021, doi: 10.1007/s10579-020-09502-8.

[34] Z. Waseem, "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter," pp. 138–142, 2016, doi: 10.18653/v1/w16-5618.

[35] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," pp. 88–93, 2016, doi: 10.18653/v1/n16-2013.

[36] B. Vidgen and L. Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, vol. 15, no. 12 December. 2021. doi: 10.1371/journal.pone.0243300.

[37] Z. Ziqi, D. Robinson, and T. Jonathan, "Hate Speech Detection Using a Convolution-LSTM Based Deep Neural Network," IJCCS (Indonesian J.

Comput. Cybern. Syst., vol. 11816 LNAI, no. 1, pp. 2546–2553, 2019, [Online]. Available: https://doi.org/10.475/123_4.

[38] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," IEEE Access, vol. 6, pp. 13825–13835, 2018, doi: 10.1109/ACCESS.2018.2806394.

[39] N. Vashistha and A. Zubiaga, "Online multilingual hate speech detection: Experimenting with hindi and english social media," Inf., vol. 12, no. 1, pp. 1–16, 2021, doi: 10.3390/info12010005.

[40] S. Masud et al., "Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on twitter," Proc. - Int. Conf. Data Eng., vol. 2021-April, pp. 504–515, 2021, doi: 10.1109/ICDE51399.2021.00050.

[41] C. M. V. de Andrade and M. A. Gonçalves, "Profiling Hate Speech Spreaders on Twitter: Exploiting textual analysis of tweets and combinations of multiple textual representations," CEUR Workshop Proc., vol. 2936, pp. 2186–2192, 2021.

[42] E. Ombui, L. Muchemi, and P. Wagacha, "Hate Speech Detection in Code-switched Text Messages," 3rd Int. Symp. Multidiscip. Stud. Innov. Technol. ISMSIT 2019 - Proc., pp. 1–6, 2019, doi: 10.1109/ISMSIT.2019.8932845.

[43] A. Joulin et al., "Deep Learning for Hate Speech Detection in Tweets," in Proceedings of the 26th International Conference on World Wide Web Companion, 2017, vol. 2, no. 2, pp. 759–760. doi: 10.1145/3041021.3054223.

[44] N. A. Setyadi, M. Nasrun, and C. Setianingsih, "Text Analysis for Hate Speech Detection Using Backpropagation Neural Network," Proc. - 2018 Int. Conf. Control. Electron. Renew. Energy Commun. ICCEREC 2018, pp. 159–165, 2018, doi: 10.1109/ICCEREC.2018.8712109.

[45] H. Faris, I. Aljarah, M. Habib, and P. A. Castillo, "Hate speech detection using word embedding and deep learning in the Arabic language context," ICPRAM 2020 - Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods, no. January, pp. 453–460, 2020, doi: 10.5220/0008954004530460.

[46] N. Bauwelinck, G. Jacobs, V. Hoste, and E. Lefever, "LT3 at SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (hatEval)," pp. 436–440, 2019, doi: 10.18653/v1/s19-2077.

[47] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," in Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020, pp. 1459–1467,. doi: 10.1109/WACV45572.2020.9093414.

[48] B. Vidgen and T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media," J. Inf. Technol. Polit., vol. 17, no. 1, pp. 66–78, Jan. 2020, doi: 10.1080/19331681.2019.1702607.

[49] P. Burnap and M. L. Williams, "Us and them: identifying cyber hate on Twitter across multiple protected characteristics," EPJ Data Sci., vol. 5, no. 1, 2016, doi: 10.1140/epjds/s13688-016-0072-6.

[50] B. Gambäck and U. K. Sikdar, "Using Convolutional Neural Networks to Classify Hate-Speech," no. 7491, pp. 85–90, 2017, doi: 10.18653/v1/w17-3013.

[51] S. Modha, P. Majumder, and D. Patel, "DA-LD-Hildesheim at SemEval-2019 Task 6: Tracking Offensive Content with Deep Learning using Shallow Representation," pp. 577–581, 2019, doi: 10.18653/v1/s19-2103.

[52] G. Wiedemann, E. Ruppert, and C. Biemann, "UHH-LT at SemEval-2019 Task 6: Supervised vs. Unsupervised Transfer Learning for Offensive Language Detection," pp. 782–787, 2019, doi: 10.18653/v1/s19-2137.

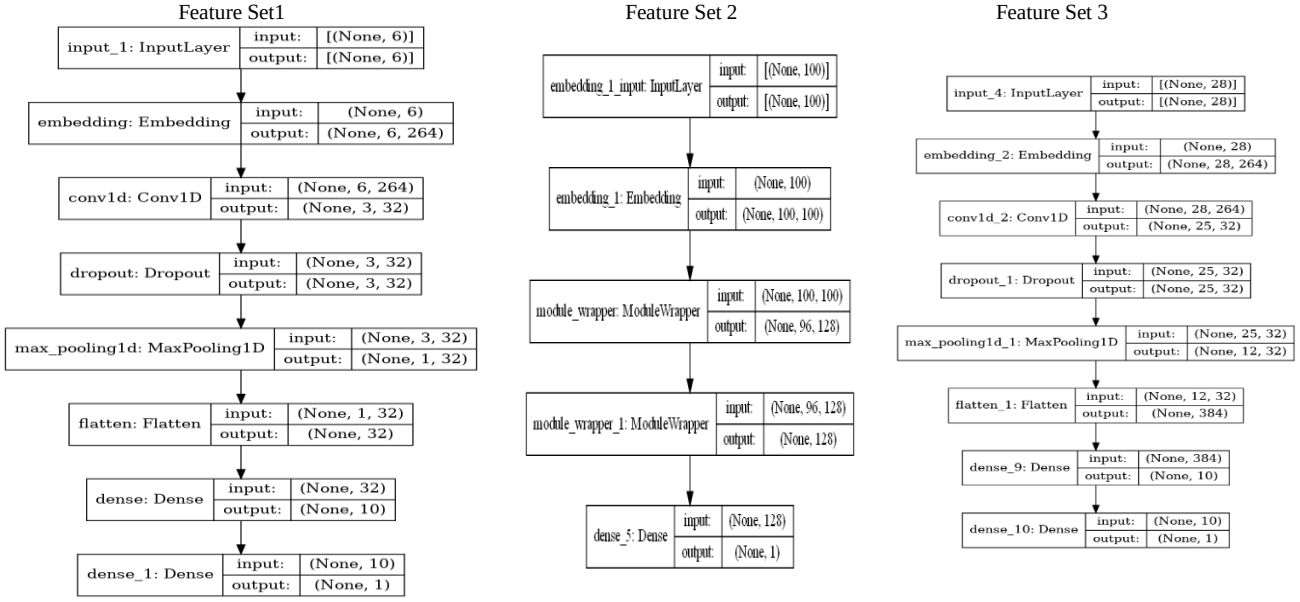
[53] K. A. Qureshi and M. Sabih, "Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text," IEEE Access, vol. 9, pp. 109465–109477, 2021, doi: 10.1109/ACCESS.2021.3101977.

[54] A. T. E. Capozzi et al., "Computational linguistics against hate: Hate speech detection and visualization on social media in the 'Contro L'Odio' project," CEUR Workshop Proc., vol. 2481, pp. 0–5, 2019.

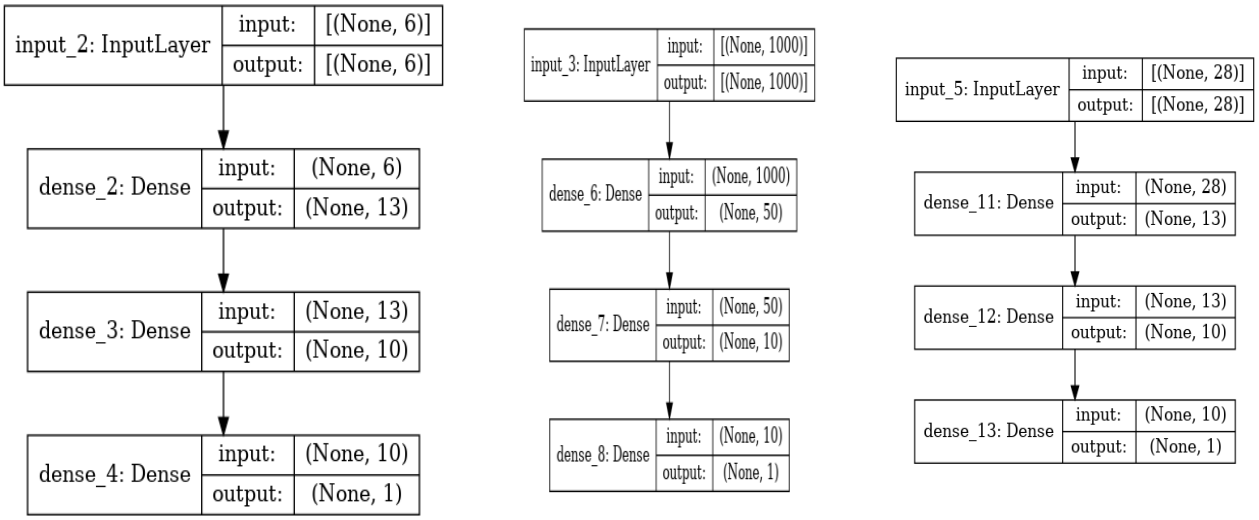
APPENDIX

Appendix A: Architecture of deep learning models used in this paper. Please note that the parameters of these algorithms was tuned based on 10% of the dataset after several trails on the algorithms till the parameters were set as shown the appendix.

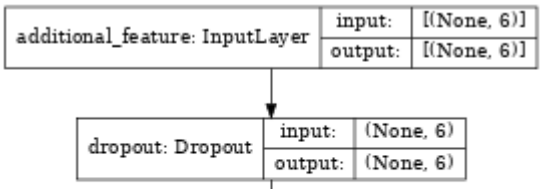
CNN



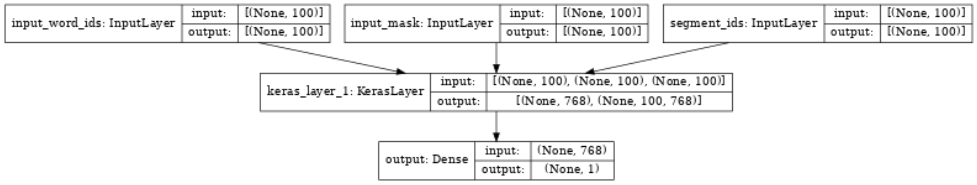
LSTM



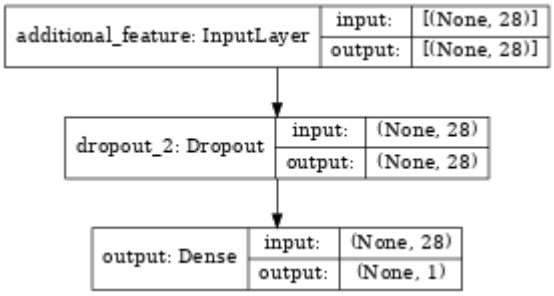
BERT
Feature set 1



BERT
Feature set 2



BERT
Feature set 3



Straggler Mitigation in Hadoop MapReduce Framework: A Review

Lukuman Saheed Ajibade[1], Kamalrulnizam Abu Bakar[2]
School of Computing
Universiti Teknologi Malaysia
Johor Bahru, Malaysia

Ahmed Aliyu[3]
Dept. of Mathematics
Bauchi State University Gadau
Nigeria

Tasneem Danish[4]
Systems and Computer Engineering Dept.
Carleton University Canada

***Abstract*—Processing huge and complex data to obtain useful information is challenging, even though several big data processing frameworks have been proposed and further enhanced. One of the prominent big data processing frameworks is MapReduce. The main concept of MapReduce framework relies on distributed and parallel processing. However, MapReduce framework is facing serious performance degradations due to the slow execution of certain tasks type called stragglers. Failing to handle stragglers causes delay and affects the overall job execution time. Meanwhile, several straggler reduction techniques have been proposed to improve the MapReduce performance. This study provides a comprehensive and qualitative review of the different existing straggler mitigation solutions. In addition, a taxonomy of the available straggler mitigation solutions is presented. Critical research issues and future research directions are identified and discussed to guide researchers and scholars.**

Keywords*—Big data; blacklisting execution; Hadoop; MapReduce; spark; speculative execution; straggler

I. Introduction

Due to the accelerated expansion of structured and unstructured data generated by the internet of things (IoT), social media, multimedia, etc., it is becoming increasingly difficult to analyse the information and data that is being generated. Applications like MapReduce, a fault-tolerant, scalable, and user-friendly framework for data processing, allow their users to efficiently process these enormous volumes of data [1], [2] . The preparation and generation of a large amount of data can be accomplished using the MapReduce approach. This is because it provides a user-friendly environment and provides solutions for a variety of ad hoc and misses, including data sorting and web indexing, among others. Bigger businesses, including Yahoo and Google, among others, use MapReduce in their large information applications.

The variety in accessibility in the CPU, I/O conflict, or network traffic is what causes stragglers. The MapReduce Framework is complete once map and reduce have been finished [3], [4]. The job is not finished in the MapReduce framework until the very reduce and map tasks are finished. Additionally, when the range of time employment increases [5-8], the number of stragglers decreases. Some compute nodes are quicker than others in a diverse environment. Faster compute nodes will finish their work ahead of schedule and wait for the stragglers to complete. Slower compute nodes are known as stragglers (Fig. 1). Nodes can occasionally fail owing to hardware or software issues. To prevent system performance degradation, it is crucial to identify stragglers at an early stage.

Traditional database management solutions such as E-R model are no longer suitable for processing and analysing of massive amounts of data generated by today's big enterprises. The bulk processing problem has become a major difficulty, and its analytical tools are evolving quickly because of Google's creation of MapReduce, which enabled millions of users to locate material from millions of pages in less than one tenth of a second. On the other hand, stragglers are widely acknowledged as a significant bottleneck in the processing of large amounts of data and they can have a considerable effect on it. Some stragglers mitigation techniques are evaluated in this paper.

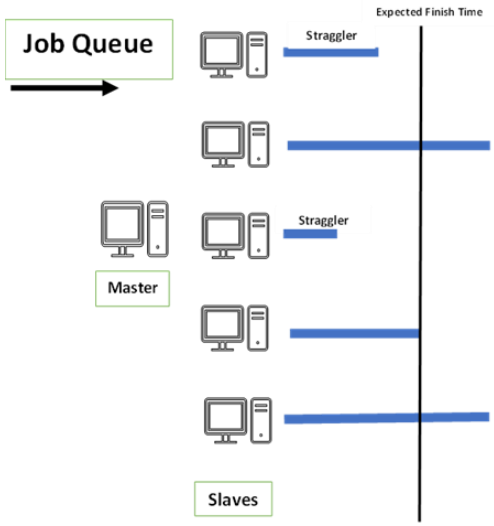


Fig. 1. Straggler Nodes in Parallel Processing.

II. MapReduce Framework and Stragglers

For significant information preparation on bunch-based figure designs, MapReduce is the ideal matching information preparing model that has been suggested [5]. Inside server centres, this system is utilised to support machine learning, data mining, and search applications. Large-scale online search applications must be addressed by the philosophy. Google was the one that originally recommended handling extremely large-scale online search applications.

Programmers are granted licences to extricate themselves from issues like parallelization, booking, and allocating so they may concentrate on creating applications. Processing, storing, visualising, and interpreting big data are the four main components of contemporary businesses and organisations. Applications on a parallel hardware cluster can be automatically run by MapReduce. Terabytes and petabytes of data can also be processed more quickly.

Due to the MapReduce capability to offer a highly effective and efficient framework for the parallel execution of applications, data allocation in distributed database systems, and fault tolerance network connections, it has recently grown in popularity in a variety of applications. Parallel map assignments, as shown in Fig. 1, are carried out as a single input data set consisting of a collection of "key value" sets that are further divided into *fixed produce* and *size blocks* transitional output. Information preparation tools called *map* and *reduce* are included in the MapReduce programming model as depicted in Fig. 2.

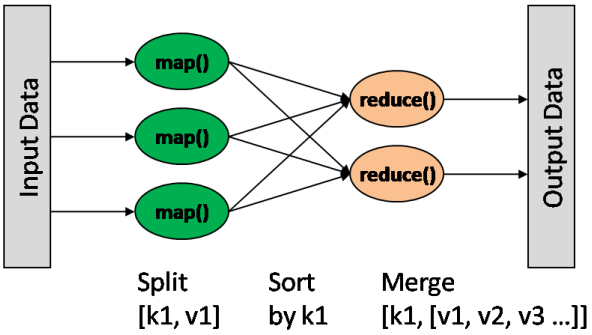


Fig. 2. MapReduce Phases.

When a user submits a job request during the Map-phase, the tasks are mapped to commodity machines for execution. In the Reduce phase, the Combiner lowers the network's data transmission rate. The Reduce-phase includes the Sort or Merging step. The time is utilised to combine the Map outputs from various nodes, and this combining is referred to as the Reduce time. The final stage in running the operation in a MapReduce fashion is the Reduce-part. The impact of each step in this process on runtime varies, so different weights should be used to estimate each job's completion time (the impact of each step on the execution process is determined by the ratio of the time of each step to the overall process's runtime).

The tasks that require more time to complete than comparable tasks are known as stragglers. There are many reasons for delaying the assignment, including the use of inefficient machines, the amount of information to process, framework obstructions, equipment heterogeneity, and competition for the available resources [6], [7].

Additionally, if one task is running slowly on a particular system, it is not important for other upcoming and current tasks to execute slowly on that same machine. Three main mechanisms must be kept in mind when addressing the straggler issue:

- If it is discovered that the anticipated remaining time is longer than the typical runtime, the procedure may be resumed up to three times.
- A speculative duplicate is scheduled if the resource measurement lowers unfavourably. The following procedures estimate the expected remaining time (trem) and the typical runtime (tnew), as shown.
- term = (telapsed * d/dread) + twrapup.
- tnew = processRate∗locationFactor∗d+schedLag.

Stragglers can occur for a variety of reasons, such as load inequality, ineffective scheduling, data localization, communication overheads, and hardware heterogeneity [8], [9]. Additionally, there have been initiatives to address one or more of these worries to lessen the issue [10]–[12]. Even if all these earlier attempts were significant and helpful in solving this issue, more rigorous analytical techniques are required to fully comprehend the effects of stragglers on the performance slowdown in huge data [13], [14].

III. Related Works

For addressing data skew for joins in a MapReduce system and avoiding stragglers, the SharesSkew algorithm was suggested by [7]. When data is skewed, the method determines the multi-way join in MapReduce. In essence, the method divides up the work of performing multi-way joins and maximises the amount of information transferred from the Mappers to the Reducers.

The technique uses a modified version of the SharesSkew algorithm to partition and share highly valued records in a distinctive way to minimise communication costs. The algorithm determines the heavy hitter value of an attribute based on the sizes of the relations or the portion of the connection with heavy hitters and how the sizes interact with one another.

In contrast to existing techniques that limit the number of Reducers employed, the SharesSkew approach merely limits the number of tuples of each Reducer. As a result, the number of tuples selected ensures that the data is distributed equally among the Reducers. (For determining the parameters of the proposed approach, both chain and symmetric joins are taken into consideration.).

A dynamic skew mitigation approach called SkewTune in MapReduce applications was proposed by [15]. The SkewTune approach tries to address the following challenges: i) the MapReduce system should not require extra input from user ii) the system should be fully transparent and iii) there should be minimal overhead even when there is no skew. If the node in the cluster is idle, the SkewTune recognizes the task with the highest anticipated remaining processing time. Afterwards, the non-processed input data of the straggling task is proactively re-partitioned such that it fully utilizes the nodes within cluster. It then conserves the ordering of the input data for the original output to be re-built by concatenation. The SkewTune is implemented as an extension to Hadoop and the efficiency is evaluated by employing several actual applications.

In the quest to address the problem of load imbalance due to data skew, a load balancing based on join algorithms in MapReduce systems was proposed by [16]. The load balancing algorithm named Fine-Grained partitioning for Skew Data (FGSD) for reduced tasks. The FGSD employs the properties of both output and input data via a proposed stream sampling algorithm. In addition, FGSD provides an approach that distributes the input data that help in handling efficient redistribution and join product skew.

Consequently, the authors declare that FGSD achieved better balancing of data distribution and minimizes execution time of jobs with different degrees of data skew. FGSD does not need any alteration to the MapReduce configuration and is suitable for handling complex jobs. Similarly, Gavagsaz et al. [17] focus on reducer phase to achieve load balancing in MapReduce system by employing scalable straightforward random sampling. The major problem in reducer phase is data skew, which lead to a significant load imbalance and degradation of performance. Therefore, a sorted balance algorithm was proposed, which is centered on sampling results. The Sorted Balance algorithm using SCalable random sampling (SBaSC). The scalable sampling algorithm is employed for monitoring a more exact approximate distribution of the keys through sampling small fraction of the intermediate data.

In MapReduce, reducer side data skew occurs due to unbalanced allocation of intermediate map-output to reducers. Therefore, [18] proposed an adaptive Learning Automata Hash Partitioning (LAHP) algorithm to address the data skew problem. The LAHP is based on learning automata game for conventional allocation of intermediate key-value pairs to designated reducers. It is achieved by setting a learning automaton on each mapper node to control the allocated load on each reducer. Thus, during execution of job, a learning automata game is enabled.

In addition, the LAHP algorithm partitions the intermediate key value pairs arbitrarily without considering the statistical distribution of pre-processing and input data. A load balancing mechanism that enhances MapReduce in Hadoop was proposed for mitigating negative impact of data skew on the performance of MapReduce [19]. Data skew has become a typical problem in MapReduce processing for handling data intensive applications. The mechanisms integrate Reservoir Sampling and Greedy (RSG) algorithms. It further slots in the concept of data locality in order to properly distribute the workload of each reducer, which is based on priority-based load-balancing mechanism (PLBM).

Wang *et al.* [20] proposed an enhanced Replication Framework of Stragglers over a Large-scale Parallel processing (RFSLP) for addressing the latency Framework of Stragglers encountered due to replication of stragglers. The framework analyzes replication latency-cost tradeoff and determines the best replication strategy. The strategy considers three design ideas including i) how many replicas are required ii) the time to replicate straggling tasks and iii) whether to terminate main copy or not. The framework analysis demonstrates that for specific execution time allocation, a small quantity of task replication can drastically minimize the cost of computing resources and latency. Further, an algorithm that estimates cost and latency based on the empirical allocation of task execution period.

In another aspect, a Framework for Assessing Stragglers Detection (FASD) mechanisms over MapReduce was proposed by [21]. It focuses on detection of stragglers because most of the existing works are centered on mitigating stragglers. In this light, an all-inclusive framework for straggler detection and mitigation was proposed. The detection strategy considers set of metrics that can be employed for characterizing and detecting stragglers. The metrics include fake positive, recall, detection latency, precision, and undetected time. Further, an architectural model was developed in such a way that the metrics can be linked to determine performance. The performance measure includes system energy overhead and execution time. To demonstrate those metrics that are effective in detecting stragglers and predict effectiveness in terms of energy efficiency performance, a number of experiments were conducted.

Similarly, a data partitioning concept, which is based on intermediate node for mitigating skew over a spark computing environment was proposed [22]. The main issue targeted in this work is unbalanced partitioning, which leads to variation in the amount of data processed by each Reducer task. Considering the mentioned issue, a Spark Key Reassigning and Splitting Partitioning (SKRSP) algorithm for handling the partition skew from the source codes of Spark-core 2.11 has been developed. The concept considers two approaches of balancing namely: partition balance for intermediate data and partition balance after shuffle operators. The contribution is in two folds first, a Key Reassigning Harsh-based Partitioning (KRHP) and rang-based Key Splitting Reassigning Partitioning (KSRP) algorithms. These algorithms can create suitable strategy for implementing the skew in the shuffle phase. The KSRP creates a weighted bound for partitioning intermediate data for the kind of sort-based applications. While KRHP stores these reassigned keys, and the new reducers of these keys are from other applications.

A proactive method named Hummer-1 for mitigating stragglers based on partial clones was proposed by [23]. In the existing solutions, different methods have been suggested including speculative execution, blacklisting and proactive mitigation. However, these solutions either waste much resource or consume much time during execution. The Hummer method trigger clones just when jobs are submitted thus, tasks in one job are assigned with clones to reduce stragglers. The initial default number of clones for a single task is three, which has been found to be the best value since there exist variations among nodes in the cluster [23]. To further improve Hummer-1, Hummer-2 was introduced which uses cloning for only tasks with high-risk delay. The authors claim that the Hummer method consumed fewer resources and minimize job delay that is, job completion time is reduced.

A Dynamic Server Blacklisting (DSB) framework was proposed for lessening stragglers to evade Quality of Service violation for time-sensitive applications [24]. The straggler task may occur due to one or more of the following reasons: heterogeneous hardware configuration, resource contention and

so on. Straggler task could become severe due to increased complexity and system scale.

The DSB is developed based on the two prominent concepts namely speculative execution, which is automatic/dynamic and blacklisting, which is manual configuration. DSB considers the previous, which is historical and present behavior of server node to improve straggler mitigation efficiency. The computing servers are ranked at a given time interval according to their present performance in completing jobs instead of their physical facilities. The servers with worst performance are momentarily blacklisted dynamically. Thus, due to the strategy no new replications/tasks are allotted to those straggler-prone nodes in each time window. DSB further offers an alternative API in such a way that the top worst nodes are blacklisted based on their ranking. An optimal node is examined as a trade-off between straggler mitigation efficiency and capacity loss. Table I contains the comparison of existing load balancing solutions.

TABLE I. COMPARISON OF LOAD BALANCING SOLUTIONS FOR STRAGGLER MITIGATION

Existing Solutions	Comparisons						Remark
	Heterogeneity among Network Node	Data Processing Method	Priority-based Scheduling	Mitigatio0n Approaches			
				Speculative	Blacklisting	Proactive	
ShareSkew [7]	No	MapReduce/Hadoop	No	Yes	No	No	Multi-way joins have not been considered to investigate an efficient multi-round MapReduce algorithm.
SkewTune [8]	Yes	MapReduce/Hadoop	No	Yes	No	Yes	SkewTune approach may lead to resource contention due to high computation.
FGSD [9]	No	MapReduce/Spark	No	Yes	No	No	Similarity joins has not been considered in the Fine-grained skew data method
SBaSC [10]	No	MapReduce/Spark	No	Yes	No	No	The query level load balancing and fairness need to be optimized
LAHP [11]	Yes	MapReduce/Hadoop	No	Yes	No	No	Data skew could also occur when the sizes of the keys are different and affect the shuffle time.
PLBM [12]	No	MapReduce/Hadoop	Yes	Yes	No	No	The transfer cost is only based on splitting capability. Which may not be sufficient achieving efficient load balancing.
RFSLP [13]	No	MapReduce/Spark	No	Yes	No	No	Despite the use of scheduling concept, priority has not been assigned some critical tasks.
FASD [14]	Yes	MapReduce/Hadoop	No	Yes	No	No	The metric evaluation has limitation of not able to detect stragglers before occurrence.
SKRSP [15]	No	Hadoop/Spark	No	Yes	No	No	In the intermediate data distribution, the weight of each key is calculated, which could lead to overhead that may result in data processing delay
Hummer-1 [16]	Yes	MapReduce	No	No	No	Yes	Computation time required for deciding which task need to be cloned could also increase the execution time
Hummer-2 [16]	Yes	MapReduce	No	No	No	Yes	The approach could lead to increase in execution time
DSB [17]	Yes	MapReduce/Hadoop	No	Yes	Yes	No	Stragglers due to data skew have not been considered in this framework.
SkewTune [8]	Yes	MapReduce/Hadoop	No	Yes	No	Yes	SkewTune approach may lead to resource contention due to high computation.

IV. Scheduling in Straggler Mitigation

In this subsection, the solutions that employ scheduling concepts considering adaptive, resource allocation and data locality-aware scheduling in MapReduce framework have been analyzed and presented.

A. Adaptive Scheduling

A problem of omission failure due to stragglers has been addressed by proposing a Failure Detector Abstraction (FDA) based on MapReduce system [25]. The omission failure is due to timeout service adjustment, which strongly endangers the workload performance. Various algorithms have been suggested based on detector abstraction for describing the timeout. Therefore, three different levels of failure detector abstractions have been suggested namely, High Relax Failure Detector (HR-FD), Medium Relax Failure Detector (MR-FD) and Low Relax Failure Detector (LR-FD). The HR-FD serves as a non-dynamic alternative to the default timeout. The MRFD acts as non-static detector that modifies the timeout, based on progress score of each workload. While the LR-FD merges the MapReduce, non-static timeout using an exterior monitoring system to enforce accurate failure detection. The LR-FD is considering in case if there are strict deadline bounded user requests. Meanwhile, the authors claim that there is significant improvement in the timeout selection for user request regardless of the failure injection time and workload type. A Task scheduling optimization framework named ET-scheduler was proposed to handle time sensitive jobs and high resource consumption [26]. The existing scheduling technique cannot complete job within the time constraint of the user. Therefore, the ET-scheduler tries to allocate resources to the tasks of job submitted. The scheduler makes sure that jobs are completed within the time specified by user. It minimizes consumption and modifies the time allocation in the process of Map and Reduce.

A Map-Balance-Reduce (MBR) programming model was proposed for improving parallel programming model for load balancing over MapReduce [27]. The problem of load imbalance occurs if the data matching to a specific key or several keys account for majority of the data, then the Reduce node task will create unbalanced load. The MBR model runs on the custom Hadoop framework, which effectively processed the unique data with unbalance data. MBR programming model is designed based on two varied scheduling namely, processing and self-adaptation scheduling. The processing scheduling in MBR tries to find unbalanced task in advance, to compile balance function. Then value/keys are pairs outputted by Map, which are transmitted to balance the function. The values are outputted by Map and can be pre-processed for unbalanced data by calling the balance function process. In the self-adaptation scheduling, if there exists unbalanced load, the present Reduce task is terminated and then the unbalanced load is dynamically split and schedule for distribution to attain dynamic load balancing of the requested task.

Cheng *et al.* [28] proposed an enhanced MapReduce solution using Adaptive task tuning (*Ant*) over a heterogeneous environment. The solution tries to address poor performance due to heterogeneous clusters. In the existing work, there is homogeneous configuration of tasks on heterogeneous nodes, which leads to load imbalance and thus causes poor performance. The *Ant* can automatically determine the optimal configuration for distinctive tasks executed on different nodes. *Ant* algorithm performs better even when the jobs are large with more than one rounds of map task execution. At the beginning task are configured with randomly chosen settings. To evade trapping in local optima and speed-up task tuning, the algorithm employs genetic functions during task configuration.

B. Resource Allocation Scheduling

Huang *et al.* [29] proposed a Workload Alleviation Scheduling Framework (WASF) in order to avoid negative effect of intermediate data skew in small scale over MapReduce cloud. The intermediate data skew is caused due to unevenly allocation of intermediate data between nodes at run time. Thus, the intermediate data skew makes the nodes in the MapReduce cloud idle, which in turn leads to waste of computation resources. This also leads to prolonging of execution time, which gives user a bad experience in cloud computing. The WASF dynamically and smartly used the available computation resources for minimizing the intermediate data skew. A method that employs result analysis of profiling and relation of system parameters was proposed to address the limitation of speculative and clone execution method [30]. The limitation is in terms of performance reduction due to heterogeneous clusters and task stragglers in big data processing. The method tunes the quantity of task slots of nodes dynamically to match the processing capability of the nodes, which is based on present task progress rate and resource consumption. Therefore, a Task Progress Rate-based (TPR) approach has been developed. The tuning process is further optimized to achieve faster convergence. Thus, the method is implemented in the Hadoop MapReduce platform.

In [11], a Root-cause analysis for stragglers in Big data environment named BigRoots was proposed for handling user programs optimization problem. The BigRoots is a general method that incorporates system features and framework for root-cause analysis of stragglers in big data environment. It analyses the stragglers using features from Big data framework including system resource utilization, JVM garbage collection time and shuffle read/write bytes. The system resources include input/output, central processing unit and network, which can detect both external and internal causes of stragglers. The BigRoots is evaluated by injecting high resource utilization over dissimilar system components and different case studies were considered to analyze dissimilar workloads in Hibench to evaluate the performance.

Lakshmi [31] proposed an algorithm for enhancing Map and Shuffle Phases (MSP) over Hadoop MapReduce in Big data environment. In the MapReduce, the shuffle phase uses individual shuffle services component with efficient input/output policy. Meanwhile, the map phase's output serves as an input to the subsequent phase. Thus, map phase requires intermediate checkpoints that regularly observe all splits created by intermediate phase. Therefore, the algorithm is designed as shuffle as a service component for decreasing the total execution time of task, monitoring map phase based on skew handling and improve resource consumption in a cluster.

C. Data Locality-aware Scheduling

A MapReduce concept based on data routing and locality was proposed to handle data imbalance in local and remote machines and to avoid network congestion [32]. A scheduling and routing algorithm named Joint Scheduler was proposed to balance task allocation to local and remote machines and to provide data routing that evade network congestion. The proposed algorithm is centered on bringing data close to computation instead of bringing computation close to data. Hence, it uses both communication network and computing resources efficiently. It is proven that the Joint Scheduler can support any load of jobs as used in the existing algorithm, which achieves the highest capacity region.

In [33], a task scheduling algorithm named rTuner was proposed to improve performance of the MapReduce job. The existing solutions are faced with the limitation of heterogeneity and resource contention, which lead to performance degradation in terms of overall job execution time. Thus, the rTuner consider the key objective to improve the reduce task execution time in both heterogeneous and homogeneous settings. Unlike the map task, the reduce task involves three phases namely, copy, shuffle and reduce phases. If the underlying situation is not analyzed by the scheduling algorithm, re-scheduling a straggler reduce task might negatively impact on the performance of the system.

Therefore, the rTuner study the reduce tasks' straggling causes and then tunes the reduce task. If tasks happen to be a straggler, then the rTuner re-schedules it to a suitable node, which depends on the situation. In summary, Table II presents the comparison of existing scheduling solutions in straggler mitigation.

TABLE II. COMPARISON OF SCHEDULING SOLUTIONS FOR STRAGGLER MITIGATION

Existing Solutions	Comparisons						Remark
	Heterogeneity among Network Node	Data Processing Method	Priority-based Scheduling	Mitigatio0n Approaches			
				Speculative	Blacklisting	Proactive	
FDA [19]	Yes	MapReduce	Yes	Yes	No	No	The failure detector abstraction did not consider data intensive computing systems.
ET-scheduler [20]	No	MapReduce/Hadoop	No	Yes	No	No	The scheduling optimization does not consider prioritization of task
MBR [21]	No	MapReduce/Hadoop	No	Yes	No	No	The pre-processing scheduling of the MBR model does not consider prioritization of the critical task
Ant [22]	Yes	MapReduce	No	Yes	No	No	The multi-tenant MapReduce settings has not been considered
WASF [23]	No	MapReduce/Hadoop	No	Yes	No	No	The scheduling framework does not consider prioritizing smaller or bigger workload
TPR [24]	Yes	MapReduce/Hadoop	Yes	Yes	No	No	This approach does not consider the shuffle phase.
BigRoots [25]	No	Spark/MapReduce	No	Yes	No	No	The relationship between locality and network utilization has not been investigated for the Root cause
MSP [26]	No	Hadoop/MadReduce	Yes	Yes	No	No	Meanwhile, heterogeneity of the network nodes has not been considered, which is important in the case of resource contention
Joint Scheduler [26]	No	MapReduce	Yes	No	No	Yes	Heterogeneity among network nodes has not been considered.
rTuner [27]	Yes	MapReduce/Hadoop	No	Yes	No	No	It is often fuzzy when deciding to declare a task as a straggler. However, the fuzziness has not been considered in rTuner.

Advantages and Disadvantages of Load Balancing Techniques (Algorithms).

Load techniques are either static or dynamic and each one has its own limitations and advantages which includes:

Advantages

- The static load balancing techniques are usually very efficient in stable environment because they do not need to monitor the resources during run-time.
- In a stable environment, operational properties do not change over time and loads are generally uniform and constant at the running time.
- Dynamic load balancing techniques are more flexible in dynamic computing environments.
- Dynamic load balancing techniques usually take into consideration different types of attributes in the environment both prior to and during run-time.
- Dynamic techniques can consider changes and provide better results in heterogeneous and dynamic environments.

Disadvantages

- Static load balancing techniques are not flexible and cannot accept changes of attributes during execution time.
- Static load balancing techniques do not consider continuous monitoring of the nodes hence they cannot consider load changes during run-time.
- When dynamic load balancing considers all changes during runtime it become more complex and dynamic to handle.
- Under certain conditions, dynamic load balancing techniques tend to have decreased performance in services.

V. Open Issues and Research Challenges

In this section, we have highlighted many research issues, which need research attention to attain efficient and effective straggler mitigation in MapReduce framework. The research issues are focused on how to balance the distribution of loads across the machines and how to efficiently schedule tasks to resource of the machines in order minimize slow tasks, which causes delay and negatively affect job completion time. The detailed discussions of the issues are as follows:

- Data Skew Caused by Inefficient Distribution of Data in Reducer Phase.

The main issues that affect the performance of the MapReduce framework is that some task take longer execution time to finish than others. This is due to data skew. The data skew is termed as inequality in the quantity of data allocated to each task or imbalance in the amount of work needed to process such data. These kinds of data are usually skewed in nature. Thus, it causes poor parallel processing, inequality of reducers input and high varied reducer execution time hence, it enlarges the completion time of the MapReduce job. Further, the intermediate data sharing in input data is not known, thus generating a strategy for the data group adjustment is difficult. It leads to the imbalance in the data distribution for a given task, which in turn causes stragglers. In addition, data skew could also occur when the sizes of keys are different and affect the shuffle time, which may cause straggler. Therefore, there is a need to develop a strategy that determines the values and keys for achieving balanced distribution of data, which mitigates the skewness of the data and improve the job completion time.

- Data Replication and Placement Issue.

The MapReduce framework is well known for its ability to handle large task and perform parallel processing during task handling. These strengths have encouraged researcher to employ replication strategy to minimize latency in job completion time. However, the replication concept has caused redundancy in task execution, which causes high resource consumption. Since the replication strategy generate redundant data there is a need for concepts that consider queuing and priority of the replicated data in terms of critical tasks for efficient job completion time. Because when there is large number of tasks that need to be executed, then the replication of these tasks will have impact on the computing resources hence, also affecting the task execution time due to resource constrain, which could cause stragglers.

- Poor Resource Allocation for Computation-Intensive Tasks.

In the existing MapReduce framework, some solutions use computational resources in a loose way on the basis that numerous idle nodes available can be used to collaboratively handle intermediate data skew. However, MapReduce system could be on a small scale and/or the task could be in a large scale. The proper utilization of the resource is very important in the case of a sophisticated system. In the existing solution, they smartly utilize computation resources in nodes and dynamically distribute workload of a node with other nodes by dispatching skewed intermediate data to a resource allocator. In addition, heterogeneity of the network nodes when allocating computation-intensive tasks to machines has not been considered, which is very significant in the case of resource contention. Consequently, the improper resource allocation to task could also lead to creation of straggler, which affect the job completion time of the Map Reduce framework. Therefore, considering the challenges, there is a need to design and develop an improved straggler mitigation solution that considers efficient resource allocation for task distribution.

- Inefficient Task-Resource Matching.

This usually results in sending simple tasks to machines with high computational capabilities and complex tasks to slow machines which may end up increasing the total job completion time.

VI. Conclusions

We have extensively reviewed existing related works and present the most recent research development in straggler mitigation approaches. The straggler issue has become challenging in MapReduce framework. Considering the negative effects of straggler, several solutions have been proposed that focus on load balancing and scheduling of the distributed task. Thus, a comprehensive review of the existing studies has been suggested in this paper. This review classified load balancing solutions into data skew and replication/placement approaches. While the scheduling approaches are classified into adaptive, resource allocation and data locality-aware scheduling. Further, open issues and research challenges are highlighted. The straggler problem degrades the performance of the existing data processing frameworks, specifically MapReduce. Therefore, there is a need to further explore more robust solution on how to effectively mitigate stragglers.

References

[1] G. Ananthanarayanan et al., "Scarlett: Coping with Skewed Content Popularity in MapReduce Clusters.".

[2] I. A. T. Hashem et al., "MapReduce scheduling algorithms: a review," Journal of Supercomputing, vol. 76, no. 7, pp. 4915–4945, Jul. 2020, doi: 10.1007/s11227-018-2719-5.

[3] G. Ananthanarayanan, A. Ghodsi, S. Shenker, and I. Stoica, "Effective Straggler Mitigation: Attack of the Clones.," Nsdi, pp. 185–198, 2013, doi: 10.1.1.366.6261.

[4] M. Reissig, "New Trends in the Theory of Nonlinear Weakly Hyperbolic Equations of Second Order," 1997.

[5] S. N. Khezr and N. J. Navimipour, "MapReduce and Its Applications, Challenges, and Architecture: a Comprehensive Review and Directions for Future Research," Journal of Grid Computing, vol. 15, no. 3. Springer Netherlands, pp. 295–321, Sep. 01, 2017. doi: 10.1007/s10723-017-9408-0.

[6] G. Ananthanarayanan et al., "Reining in the outliers in map-reduce clusters using mantri," Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2010, pp. 265–278, 2019.

[7] M. Zaharia, A. Konwinski, A. D. Joseph, and R. Katz, "Improving MapReduce Performance in Heterogeneous Environments."

[8] J. Rey, M. Cogorno, S. Nesmachnow, and L. A. Steffenel, "Efficient prototyping of fault tolerant map-reduce applications with Docker-Hadoop," in Proceedings - 2015 IEEE International Conference on Cloud Engineering, IC2E 2015, 2015, pp. 369–376. doi: 10.1109/IC2E.2015.73.

[9] U. Kumar and J. Kumar, "A Comprehensive Review of Straggler Handling Algorithms for MapReduce Framework," International Journal of Grid and Distributed Computing, vol. 7, no. 4, pp. 139–148, Aug. 2014, doi: 10.14257/ijgdc.2014.7.4.13.

[10] Y. Chen, S. Alspaugh, and R. H. Katz, "Design Insights for MapReduce from Diverse Production Workloads," 2012. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-17.html

[11] H. Zhou, Y. Li, H. Yang, J. Jia, and W. Li, "BigRoots: An Effective Approach for Root-Cause Analysis of Stragglers in Big Data System," IEEE Access, vol. 6, pp. 41966–41977, 2018, doi: 10.1109/ACCESS.2018.2859826.

[12] M. Fatih, A. Aktas¸, P. Peng, and E. Soljanin, "Effective Straggler Mitigation: Which Clones Should Attack and When?"

[13] A. Kamal Abasi, A. Tajudin Khader, M. Azmi Al-Betar, S. Naim, S. Naser Makhadmeh, and Z. Abdi Alkareem Alyasseri, "A Text Feature Selection Technique based on Binary Multi-Verse Optimizer for Text Clustering; A Text Feature Selection Technique based on Binary Multi-Verse Optimizer for Text Clustering," 2019. [Online]. Available: http://www.unine.ch/Info/clef/,

[14] A. H. Katrawi, R. Abdullah, M. Anbar, and A. K. Abasi, "Earlier stage for straggler detection and handling using combined CPU test and LATE methodology," International Journal of Electrical and Computer Engineering, vol. 10, no. 5, pp. 4910–4917, Oct. 2020, doi: 10.11591/ijece.v10i5.pp4910-4917.

[15] Y. C. Kwon, M. Balazinska, B. Howe, and J. Rolia, "SkewTune in action: Mitigating skew in MapReduce applications," Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 1934–1937, 2012, doi: 10.14778/2367502.2367541.

[16] E. Gavagsaz, A. Rezaee, and H. Haj Seyyed Javadi, "Load balancing in join algorithms for skewed data in MapReduce systems," Journal of Supercomputing, vol. 75, no. 1, pp. 228–254, 2019, doi: 10.1007/s11227-018-2578-0.

[17] E. Gavagsaz, A. Rezaee, and H. Haj Seyyed Javadi, "Load balancing in reducers for skewed data in MapReduce systems by using scalable simple random sampling," Journal of Supercomputing, vol. 74, no. 7, pp. 3415–3440, 2018, doi: 10.1007/s11227-018-2391-9.

[18] M. A. Irandoost, A. M. Rahmani, and S. Setayeshi, "Learning automata-based algorithms for MapReduce data skewness handling," Journal of Supercomputing, vol. 75, no. 10, pp. 6488–6516, 2019, doi: 10.1007/s11227-019-02855-0.

[19] F. H. Syue, V. A. Kshirsagar, and S. C. Lo, "Improving mapreduce load balancing in hadoop," ICNC-FSKD 2018 - 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, pp. 1339–1345, 2018, doi: 10.1109/FSKD.2018.8687158.

[20] G. Joshi and G. W. Wornell, "Efficient Straggler Replication in Large-Scale," vol. 4, no. 2. 2019.

[21] T. D. Phan, G. Pallez, S. Ibrahim, and P. Raghavan, "A new framework for evaluating straggler detection mechanisms in mapreduce," ACM Transactions on Modeling and Performance Evaluation of Computing Systems, vol. 4, no. 3, 2019, doi: 10.1145/3328740.

[22] Z. Tang, W. Lv, K. Li, and K. Li, "An Intermediate Data Partition Algorithm for Skew Mitigation in Spark Computing Environment," IEEE Transactions on Cloud Computing, vol. PP, no. c, p. 1, 2018, doi: 10.1109/TCC.2018.2878838.

[23] J. Li, C. Wang, D. Li, and Z. Huang, "Partial clones for stragglers in MapReduce," Communications in Computer and Information Science, vol. 503, pp. 109–116, 2015, doi: 10.1007/978-3-662-46248-5_14.

[24] X. Ouyang, C. Wang, and J. Xu, "Mitigating stragglers to avoid QoS violation for time-critical applications through dynamic server blacklisting," Future Generation Computer Systems, vol. 101, pp. 831–842, 2019, doi: 10.1016/j.future.2019.07.017.

[25] B. Memishi, M. S. Pérez, and G. Antoniu, "Failure detector abstractions for MapReduce-based systems," Inf Sci (N Y), vol. 379, pp. 112–127, 2017, doi: 10.1016/j.ins.2016.08.013.

[26] Y. Ren, H. Li, and L. Wang, "Research on MapReduce Task Scheduling Optimization," IOP Conference Series: Materials Science and Engineering, vol. 466, no. 1. 2018. doi: 10.1088/1757-899X/466/1/012016.

[27] J. Li, Y. Liu, J. Pan, P. Zhang, W. Chen, and L. Wang, Map-Balance-Reduce: An improved parallel programming model for load balancing of MapReduce, vol. 105. Elsevier B.V., 2020, pp. 993–1001. doi: 10.1016/j.future.2017.03.013.

[28] D. Cheng, J. Rao, Y. Guo, and X. Zhou, "Improving MapReduce performance in heterogeneous environments with adaptive task tuning," Proceedings of the 15th International Middleware Conference, Middleware 2014. pp. 97–108, 2014. doi: 10.1145/2663165.2666089.

[29] T. C. Huang, K. C. Chu, J. H. Lin, G. H. Huang, and C. K. Shieh, "Workload Alleviation Scheduling Framework to Alleviate Negative Performance Impact of Intermediate Data Skew in Small-Scale MapReduce Cloud," 2018 International Conference on System Science and Engineering, ICSSE 2018, pp. 1–6, 2018, doi: 10.1109/ICSSE.2018.8520003.

[30] X. Zhao, K. Kang, Y. Sun, Y. Song, M. Xu, and T. Pan, "Insight and reduction of MapReduce stragglers in heterogeneous environment," Proceedings - IEEE International Conference on Cluster Computing, ICCC, pp. 1–8, 2013, doi: 10.1109/CLUSTER.2013.6702673.

[31] J. V. N. Lakshmi, "Data analysis on big data: Improving the map and shuffle phases in Hadoop Map Reduce," International Journal of Data

Analysis Techniques and Strategies, vol. 10, no. 3. pp. 305–316, 2018. doi: 10.1504/IJDATS.2018.094130.

[32] W. Wang and L. Ying, "Data locality in MapReduce: A network perspective," 2014 52nd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2014, pp. 1110–1117, 2014, doi: 10.1109/ALLERTON.2014.7028579.

[33] R. Patgiri and R. Das, "rTuner: A performance enhancement of MapReduce job," ACM International Conference Proceeding Series. pp. 176–186, 2018. doi: 10.1145/3177457.3191710.

Blood Management System based on Blockchain Approach: A Research Solution in Vietnam

Hieu Le Van[1], Hong Khanh Vo[2], Luong Hoang Huong[3], Phuc Nguyen Trong[4], Khoa Tran Dang[5], Khiem Huynh Gia[6], Loc Van Cao Phu[7], Duy Nguyen Truong Quoc[8], Nguyen Huyen Tran[9], Huynh Trong Nghia[10], Bang Le Khanh[11], Kiet Le Tuan[12]
FPT University, Can Tho City, Viet Nam

***Abstract*— More and more new health care solutions are born based on the development of science and technology. The subjects who benefit the most, in this case, are not only patients (i.e., shorter healing time, faster recovery) but also medical staff, e.g., doctors/nurses (i.e., easy monitoring of the patient's recovery process, proposing new treatment). However, there are still products that have not found an alternative: blood and blood products. Regardless of how science and technology can affect all aspects of patient treatment as well as medical care, blood still plays an important role in the treatment method. In addition to the above, blood and blood products may only be obtained from volunteers (i.e., blood donors). The preservation process is also very difficult, and no medical facility has enough facilities to preserve them. Therefore, the current process of blood preservation and transportation is done manually and contains many potential risks (e.g., data loss, personal information collection). In addition to the above barriers, developing countries (including Vietnam) also face many difficulties due to limited facilities. It is for this reason that this paper aims at a Blockchain-based technology solution for efficient management and distribution of blood from blood products. On the one hand, the paper contributes to the limitations in the information management process of storing and transporting blood and its products in the traditional database being applied in medical facilities in the cities and provinces in the Mekong Delta (the West-South of Vietnam). On the other hand, the article offers technology-based solutions to increase transparency and reduce the fear of centralized data storage (i.e., security and privacy issues). We also implement a proof-of-concept to evaluate the feasibility of the proposed approach.**

***Keywords*—Blood donation in Vietnam; blockchain; hyperledger fabric; blood products supply chain**

I. Introduction

There is no denying in the contribution of technology in the treatment of diseases today. Specifically, it increases the patient's ability to recover after treatment and reduces the risk of human error. In parallel with those positive aspects, new diseases always appear with more dense frequency and more dangerous toxicity for patients (e.g., Covid-19). For this reason, treatment and health care procedures have changed dramatically from the same time period ten years ago. However, some products cannot be replaced in the treatment process regardless of the development of science and technology [1]. One of the prime examples of this type of product is blood. Indeed, blood and its products are important medical resources in long-term treatment as well as in emergencies [2], e.g. blood is often required for trauma victims, surgery, organ transplants, childbirth and for patients being treated for cancer, leukemia and anemia. Each unit of blood is very precious and gives hope to the patient. For example, a liter of blood can sustain the life of a premature baby for two weeks; 40 or more units of blood may be required for the survival of an accidental blood loss trauma victim, or eight platelets per day are the minimum level for the treatment regimen of blood cancer patients.

In addition to the irreplaceable requirements in the treatment of diseases, the supply of blood is extremely limited because only donated blood is used instead of other substitutes due to a number of medical reasons. Besides, the requirements on time of use are also very strict to ensure the safety of the recipient. Specifically, blood and its products cannot be stored for long periods of time. Red blood cells must be used within 42 days of collection. Meanwhile, platelets have a shelf life of five days of collection. On the other hand, checking blood group similarities between blood donors and recipients is extremely strict. In particular, all blood collected must be rigorously tested to reduce the risk of transmission of infection by blood transfusion (e.g., hepatitis B and C (HBV - HCV) or human immunodeficiency virus (HCV) (HIV)) before transmission to the recipient [3]. All three of the above reasons are the main reasons for the shortage of supply according to Chapman et al. [4]. The above article also shows that the best way to optimize the use of blood and blood products is to save the amount of blood that can be received from the donor.

In addition to the above obstacles affecting the blood supply chain management process, time requirements are also extremely important. Specifically, no one expects blood, but if it's not available when it's needed, the consequences can be deadly. While donors may tell you there's no better feeling than saving a life, only about 5% of eligible donors actually donate [5]. Therefore, maximising the amount of blood stored in the warehouse is imperative.

In addition to the above barriers, one of the main obstacles in blood collection and storage for developing countries is supply (i.e., volunteers still do not have a positive attitude towards blood donation) [6], and infrastructure and database management system (DBMS) for blood and its products' storage [7]. Vietnam is also on the list of developing countries and suffers from a shortage of supply. As far as we know, there is only one hematology hospital that supplies blood to the whole Mekong Delta. This article focuses on the second issue (i.e., infrastructure and DBMS) rather than increasing citizen blood donation behaviour-aware. Therefore, in the scope of this article, we only focus on technology solutions to improve the current management system as well as change from the centralized to the decentralized storage system. In particular, we aim to share donor data in a controlled manner; for

example, reduce the medical declaration time from the second blood donation onwards. Indeed, each user can only donate blood at least 28 days after their previous blood donation (i.e., about a month) [8]. Hence, previously stored information must be stored decentralized, and volunteers can donate blood at a different location without providing the previous information.

In addition, data sharing among the hospitals and blood donation clinics makes donor management easier. Health workers have more options in contacting volunteers for the next blood donation, thereby promoting the blood donation movement in the community. To solve the second problem, we want to verify blood and blood products to achieve transparency-proof. As storage conditions (e.g. temperature, humidity) and storage time vary depending on the blood product collected. The storage and transportation of blood and blood products from storage to hospital (or vice versa) makes it difficult to identify relevant information, including time, location, health notes, and so on. To address the above issues, many methods have proposed blockchain technology to increase the transparency and traceability of information storage. There are several benefits of the blockchain approach proved in the management system. For instance, Cash-on-Delivery (COD) exploited the smart contract to set up the contract compliance among the stakeholders [9], [10], or define the shipper participant role in the marketplace [11], [12]. Moreover, in the Healthcare system, the patient can control the permission to show the necessary information to the authorized partner (e.g., nurse, doctor) [13], [14] or define the emergency between the medical staff and patient's friend or relative [15], [16] Besides, data stored decentralized is also a plus point compared to the current traditional storage model. Where all data is shared and easily traceable to identify the source of blood and blood products, however, current models cannot fully address the requirements for storing different information about blood and its products [17], [18] for proper storage (i.e., shelf-life of usage, temperature, humidity). Furthermore, there has not been an in-depth study to assess the appropriateness of the application of advanced technologies in supply chain management w.r.t. blood and its products in Vietnam. To address these problems, this paper proposes blood and its product management process by applying blockchain technology and decentralized storage for medical facilities in Vietnam.

The contribution of this article is three-fold: i) analyzing the current management mechanism of blood and its products in the provinces and cities in the Mekong Delta (southern Vietnam); ii) proposing a solution to manage the supply chain of blood and blood products based on blockchain technology; iii) implement the proposed model based on Hyperledger Fabric platform and evaluate their feasibility.

Following this introduction, the state-of-the-art is presented in Section II to summarize blockchain-based blood supply chain system approaches as well as that system for blood and its products management in Vietnam. Then, we analyze the current blood and the supply of its products chain system and our architecture before presenting the execution algorithm in Sections III and IV, respectively. Section V focuses on the analysis and evaluation. The following section presents the discussion of this article. Finally, suggestions for future research and conclusion are made in the last section.

II. Related Work

A. Blood Supply Chain Management Systems not Applying the Blockchain Technology

Supply chain management integrates core business processes and information. These processes use a central server to handle visibility and traceability issues. The system combines a very complex process that requires synchronization of different operations, leading to randomness and supply chain risk [19], [20]. For example, Nagurney et al. [21] proposed a model to minimize costs and risks by expressing the breakdown properties of blood as supply coefficients. Armaghan and Pazani [22] proposed a blood supply chain to handle urgent requests from blood units during the Iran earthquake. The authors build a multi-level, multi-objective model to find an optimal route based on the selected routes to transport blood. The main contribution of [22] is to reduce the cost of the blood supply chain network and maximize reliability. In addition, Eskandari-Khanghahi et al. [23] have developed a model that provides a combination of integer mixed linear programming while considering location, allocation, inventory, and distribution. On the other hand, Delen et al. [24] have integrated GIS (geographic information systems) and data mining techniques to build blood supply chain processes. The main purpose of [24] work is to build an optimal blood transport model to be applied in the military environment.

One disadvantage of centralized storage in the above approaches is transparency [25]. To address this issue, Lam et al. [26], [27] demonstrated the implementation of a micro services-oriented software architecture for middleware that collects, stores, and traces data in a centralized manner in order to provide data analysis. To apply these advantages, a centralized blood donation management solution has been proposed in [28]. This approach not only reduces the amount of information collected from blood donors but also improves the efficiency of blood donation management.

B. Blood Supply Chain Management Systems based on Blockchain Technology

Trieu [17], and Nga [18] propose a cold-blooded supply chain system based on Hyperledger Fabric called BloodChain. The proposed system supports the verification of blood-related transactions from donors to recipients. Moreover, BloodChain allows displaying the necessary information during the blood donation process. Specifically, the actors in the system only receive enough information to verify information about donors as well as recipients. Similarly, Lakshminarayanan et al. [29] propose a blood supply chain management system based on Hyperledger Fabric. Similar to BloodChain, it also ensures transparency of donated blood by tracking blood units between donors and recipients. Moreover, Toyoda et al. [30] have integrated the RFIDs into the blood bags using the EPC stored in the tag. This integration helps to ensure reliability and avoid tampering by tracking products and checking their tags.

However, there are some limitations to the aforementioned solutions. For example, the verification of the system proposed in [18], [17] is incomplete due to the lack of evaluation analysis. Moreover, they ignore the mobility role of the blood/its product, which exploits the transportation company [31] or

shipper [32]. In particular, the blood and its product must be received from several mobility resources (e.g., mobile blood collection units, medical clinics) and stored at the blood bank/hematological hospital. Therefore, the role of transporter[33] is vital in blood donation environments. Furthermore, the monitoring solution proposed in [30] is limited to monitoring blood bags only, and it does not guarantee the traceability of blood components (i.e., red blood cells, platelets, white blood cells, platelets and plasma). Since different blood components have different shelf lives and storage temperatures, the order of user preference should also be considered.

III. The Blood Donation Traditional Process

A. The Current Blood Donation Process in Vietnam

To get the most unbiased view of the traditional blood donation and blood handling process, we collected information about the process in hospitals and healthcare facilities in the Mekong Delta, Vietnam. We conducted a short interview with the medical officers working at the hematology hospital in Can Tho, which supplies blood and blood products to hospitals and healthcare facilities not only in Can Tho city but also in neighbouring provinces (e.g., Vinh Long, Ben Tre, Hau Giang).

Fig. 1 presents the current blood donation process. In particular, the donors are able to donate blood through four ways, including medical clinic, mobile blood collection unit, medical facility (e.g., hospital), and hematology hospital. Except for the second blood donation method (i.e., mobile blood collection unit) which is held in public places for a short time (usually 1 day), donors can donate blood at any time at the three remaining medical facilities. For blood donation at the medical clinic and mobile blood collection unit, the collected products are transferred to the storage facility at the hematology hospital. Here, blood is separated into several components including plasma, red blood cells, white blood cells, and platelets, and then stored according to the specific conditions of the blood product (e.g., temperature, humidity, duration). For the two remaining ways of donating blood, the collected blood does not need to go through a transportation step because these medical facilities have the facilities/equipment to conduct separation and storage. Finally, the blood and its products are delivered to the hospital for recipients. All these steps are performed manually and stored locally at each location (e.g., hematology hospital, medical clinic, hospital).

Although the traditional approach is simple and easy to apply to all medical facilities because it does not require high support technology as well as easy to deploy in a practical environment. However, the above approaches face many inherent risks for systems based on centralized management. Verifying the reliability of the data is admissible to this approach. In particular, any data displayed is only taken from the data available in the database, which is provided by the central server. Moreover, the important information that affects the treatment process can be lost if the central server is hacked. This is an extremely dangerous thing for medical/healthcare organizations. Due to these dangerous risks, it is urgent to find a decentralized storage solution as well as increase the authenticity of data. Blockchain technology can fulfil both of these issues. The next sections will detail the blockchain-based management models to address the current blood management model.

B. The Process Requires Blood and its Products

This procedure outlines the basic steps for transferring blood from hematology hospitals to hospitals when blood is in short supply or in an emergency (e.g., platelet request). Fig. 2 shows the procedure for requesting blood when a specific type of blood product (e.g., platelets) is needed. New blood recipients go to hospitals or healthcare centers for treatment as a first step. We assume that the requested amount of blood/its products is a rare blood type or that the treatment site has no corresponding blood/blood product. To solve this request, the medical center/hospital sends a request to the hematology hospital to find the corresponding blood/blood product source. At this point, the medical staff will find candidates based on the previous list of donors. They filter information by matching requests received from lower-level hospitals with information available from donors in the system. As a next step, medical staff at the hematology hospital contact the selected candidates to determine if they can donate blood (with a pre-set time). After selecting potential candidates (at least two people), the medical staff conducts a health assessment and blood tests to rule out those with weak health or blood problems at the time request point. Finally, they selected a single candidate to draw blood and transferred it to the respective treatment facility.

C. Limitations of the Two Traditional Processes

As described in the Introduction (i.e., Section I), all data is stored and managed centrally. The information is easily attacked by malicious users resulting in data loss. In addition, it is difficult to evaluate the stored information because all blood/its products data is displayed from the information stored in the central server. In terms of management processes, health care systems lack linkage, i.e., only supporting the link between hematology hospitals and lower-level hospitals. Hence, it is difficult to take advantage of the available blood volumes in the system. All requests are handled single-line (i.e., send directly to the hematology hospital) when an emergency occurs. Information about donors and recipients is easily stolen.

IV. Approach

A. Blood Donation Process based on Blockchain Technology

The biggest difference between the proposed model based on blockchain technology and the traditional model is that all data and retrieval requests are stored in a distributed ledger. Specifically, Fig. 3 shows the storage process of the stakeholder who has a role in the system, i.e., medical facilities (e.g., hospital, medical clinic or hematology hospital); donors; blood products (e.g., red blood cells, white blood cells, platelets); transportation; and blood bank. All data related to blood/blood products are stored, but also all requests for data retrieval from relevant parties (e.g., healthcare workers, carriers) are stored in a log book. dispersion one. This increases transparency for the whole system. Data owners easily know which users can access their data. As for blood records, all information about donors is stored in a distributed ledger. Information about blood type, time, date, and other preservation information are all stored and processed in a decentralized form. Thereby, medical facilities can retrieve and confirm data related to the treatment process. Besides, the data of medical centers/hospitals is also very

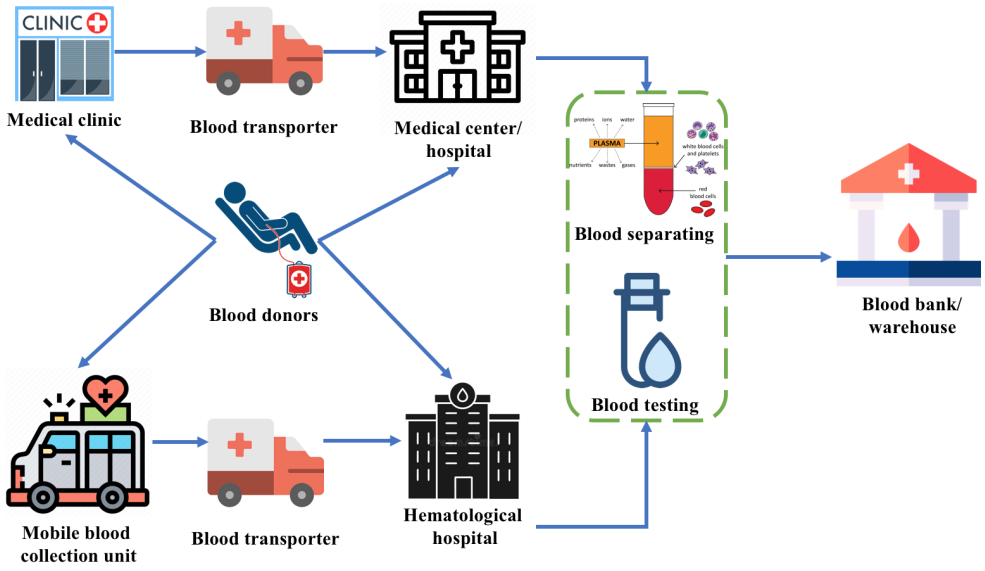


Fig. 1. The Current Blood Donation Process.

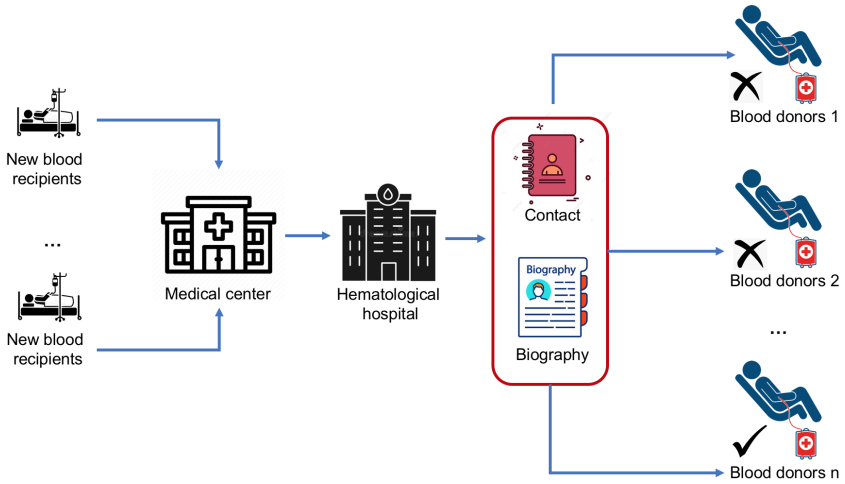


Fig. 2. The Current Blood Request Procedure

important. Instead of local storage, our proposed model is towards a decentralized model, where data can be shared for healthcare purposes. Specifically, medical centers/hospitals can exchange information on blood volume and blood products that can be shared in an emergency, reducing requirements for hematology hospitals. On the other hand, information about donors donated at one medical facility in the past can be easily retrieved by another facility, thereby increasing the quality of treatment for patients.

Fig. 4 details the process of sharing donor data between different health facilities. Specifically, basic information (i.e., biography) about addresses and phone numbers is shared for health care purposes. In addition, information about the amount of blood in stock is shared with other medical facilities. In addition, the conversion process is always up to date if there are any shipping requirements for healthcare purposes. However, the results of donor blood tests are not shared in the current model to limit privacy violations. The information shared by donors is only related to health care needs. In unsatisfactory blood test results, the donor's personal data will be deleted from the distributed ledger. We do not support off-chain executions (i.e., out-of-scope) in the current approach, so data sharers (i.e., medical officers) must secure on-chain data uploads.

B. Algorithms

In the proposed blockchain-based system, we have two main algorithms to control the process. As explained in Section IIB, our methods apply the Hyperledger Fabric platform to conduct the transactions. Moreover, due to these approaches are suitable for a hybrid business environment [32], we exploit JavaScript format for the information structure and the blockchain system for the process structure below::

```
bloodRecords = {
"donorID": donor ID,
```

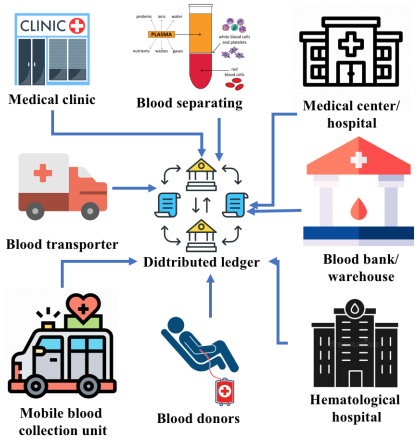


Fig. 3. The Distributed Storage of Blood Donation Process.

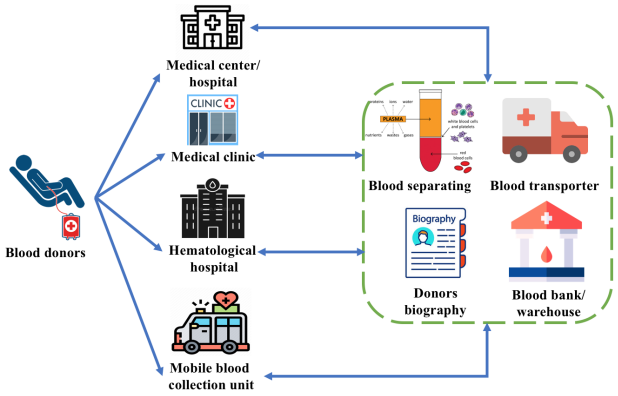


Fig. 4. The Blood Donation Process based on Blockchain Technology.

```
"bloodID": blood and its production ID,
"bloodGroup": blood group,
"bloodProduction": blood production,
"temp": temperature,
"humidity": humidity,
"time": time,
"date": date,
"duration": duration,
"state": 0,
"medicalFacility": locate of medical facility,
"amount of blood unit": 350
};
```

This section targets two main algorithms, including Algorithm 1 describing the data creation to store the blood and its products; whereas Algorithm 2 presents the delivery of blood samples from the donation place to the hematological hospital/medical facility).

Algorithm 1 present all the steps to collect the data relate to donors, blood and its product, storage requirement as well as their metadata (date, time). Blood and its products, once collected, are identified by the donorID of the respective donor. Each blood and its item has a distinct bloodID. This data is unique and not repeated in the ledger; the blood is then stored at the blood bank or the other facility of medical with the identity info being donated. Finally, all the collected data (i.e., blood and its products, their metadata) have updated the status in the ledger.

Algorithm 1 The Blood and its Products Data Creation

```
Input: Donor ID, blood group, blood production, blood donation metadata (i.e., time, date), storage requirements (i.e., temperature, humidity, duration), and blood unit amount
Output: all the related data is stored in the ledger
for blood unit and its metadata do
    storing all the blood unit and its metadata to the ledger;
end for
```

Algorithm 2 shows process of blood and its product delivery from the blood bank or donation place to the recipient (i.e., patient) or clinical center. This process may transfer among the medical facilities in emergency situations. Since the time of storage is limited, they first update the time of remaining usage and then select the oldest one to transfer the destination (i.e., other facilities or recipient). To speed up this process, the shipping company selected and detect the destination address. During the transportation process, the stakeholders are allowed to update the status of the blood units. In particular, the transportation process received several inputs namely, the shipping unit's information, blood units as well as its metadata and the output is the delivery data (time, date, address) and the updated institution of the clinical center (corresponding ID_Center).

Algorithm 2 Delivery of Blood and its Product from Donation Places (e.g., Medical Clinic) to Blood Storage (e.g., Blood Bank)

```
Input: the blood delivery unit's data
Output: the blood delivery' data
for delivery unit do
    for blood and its products do
        Blood_Unit_ID
    end for
    for blood and its products do
        update the new location of the medical center
    end for
end for
return Encrypted hash
```

V. Evaluation

A. Environmental Setting

Our proposed framework is deployed in the Hyperledger Fabric platform. We simulate the environment requirement inside docker containers. This section measures the chaincode's performance of the two scenarios in the algorithms (see Section IV-B), namely initializing and querying data. The experiments are deployed on Ubuntu 20.01 configuration, Core i5 with 2.7Ghz and 8GB RAM.

To collect all information related to the performance, we exploited the Hyperledger Caliper[1]

[1]Hyperledger Caliper deploys the test situations and gathers all the data with respect to the execution. See more related information in this link https://www.hyperledger.org/use/caliper

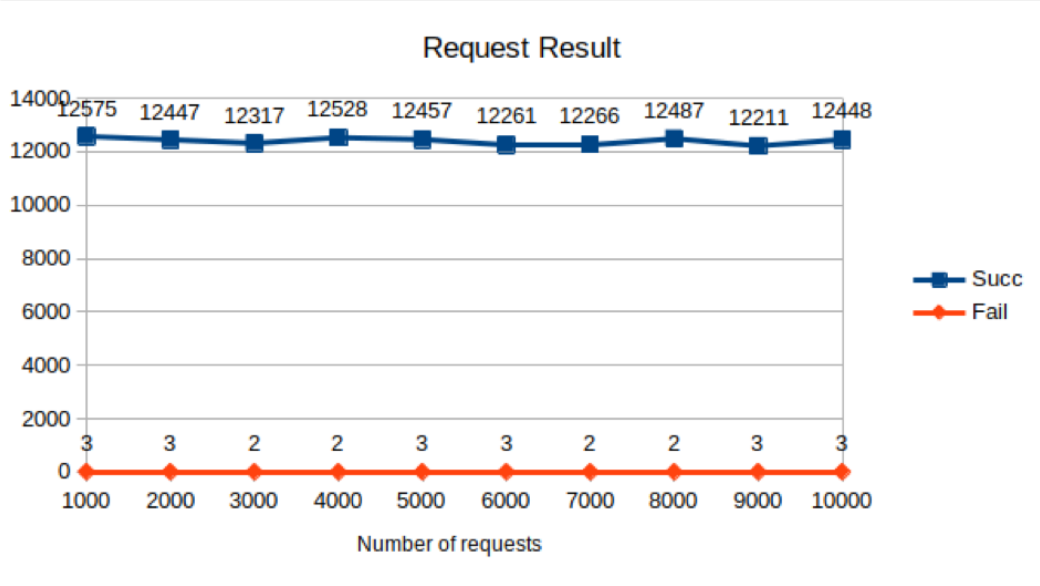


Fig. 5. The Results of the Create Data Functions.

B. Execution Time

In the data creation execution, we measure how long to response time (i.e., seconds) is to create a data request function. In this case, for each worker, they initially created 1000 requests for each second in the system. We stop at 10,000 requests for each second. The #requests are then consistently transferred to the system (i.e., at most 2 minutes of waiting) and slowly rise by 1,000 requests for each performance test until 10,000 requests/s. All the results of these executions are recorded and presented in Fig. 5. From this figure, we are able to detect that the #requests w.r.t successful execution is high compared to the other ones (especially the failed requests are tiny). In fact, the number of successful execution requests is higher than 12,000 request/s; in contrast, this amount of unsuccessful ones is a range of 2-3 requests.

Fig. 6 shows the outcomes of requests measurement that execute the update methods for blood and its products. #Workers rose by 2 in this attribute which is higher than the beginning data creation requests; the number of requests. In the same situation, we start from 1,000 requests/second and increase another 1,000 requests/second until stop at 10,000 requests/second for each worker. The number of successful execution requests increased not noticeably, range 15K to 16K requests for each second. However, compared to the initial process with only one worker, the number of failed requests is higher. To explain this point, we consider the number of workers who creates the request (i.e., the number of workers has increased) and analyse the latency aspect of the system (see more detail in the next subsection). From this point, we can claim that the time of response is longer when the number of users is increased.

For the data query, Fig. 7 describes the outcomes and quantity of requests to query the blood and its products recorded in the distributed ledger. In the last scenario, we increase the number of workers to 10 workers[2]. Similar to the two above scenarios, we start from 1,000 requests/second and increase another 1,000 requests/second until stop at 10,000 requests/second for each worker. According to the outcomes of Fig. 7, we found that the results are still steady. Specifically, #succeeded requests varies from the lowest rank, around 37K, to the highest, approximately 42K requests. Whereas #unsuccessful requests for data query are at most one request (negligible). This further proves the outstanding advantages of the distributed system compared to the traditional centralized ones.

C. Latency

In this second scenario, the article gauges and assesses the latency situation to three functions: data initialization, query, and update.

Fig. 8 indicates the latency of the information initialization function of the blood samples. In which the majority of the three latency levels are steady. Specifically, the highest latency index varies in the range of 500s when the quantity of requests rises from 1000 to 10,000 requests. However, a certain case happens in case 6 with 6000 requests/s; the latency increases significantly at 2240.21s, which may arise when the system has a bottleneck in processing transactions, so arriving transactions take more time to complete processing. Nevertheless, at Min and Avg latency indicators, the processing time for requests is still stable at the level of 363.16s/request and 0.64s/request, respectively.

The latency of data query requests is indicated in Fig. 9. The greatest latency index varies quite importantly from 1.41s/request to 4.64s/request. Nonetheless, the system's processing time progressively diminished and stayed stable when the number of requests rose from 1000 to 10,000. The most noteworthy value is 4.65s for a total of 2000 requests.

[2]We assume that workers are the doctors/nurses/officer in the medical facilities

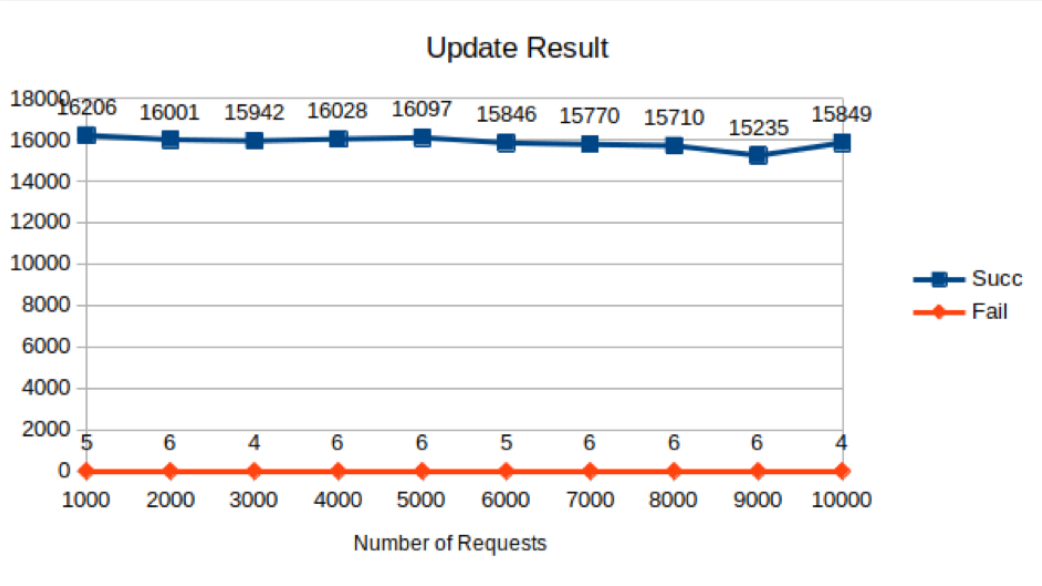


Fig. 6. The Results of the Update Data Functions.

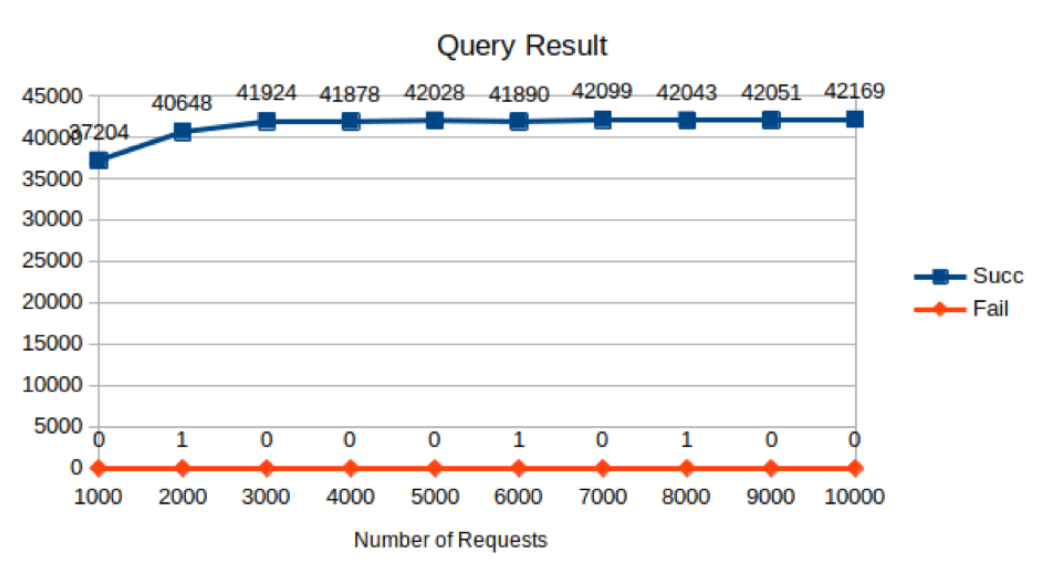


Fig. 7. The Results of the Request Data Functions.

Fig. 9 depicts the latency of requests to query information of all users in the system. The lowest latency 0.01 - 0.02 and average latency 0.19 - 0.82 are negligible even when the quantity of requests rises to 10,000 requests at the same time. The highest value of Max latency is 4.65s, correspondent with the scenario of 10 users delivering 2000 requests to the system at that point.

Ultimately, the latency of the data update function of the blood samples is also depicted in Fig. 10. The latency of the requests is kept stable when the number of requests increases from 1000 to 10,000. However, the processing time is at the most significant rank in comparison with 2 requests: to initiate and query data; this occurs when the system begins to query the information and then upgrade the data fields requesting data and stores new ones. The highest value of Max Latency is 795.4s for 1000 requests; moreover, the other cases are steady with a postponement of over 700s.

VI. Discussion

A. Remarkable

It is easy to see that the execution time of initialization, query and surrogacy requests is quite stable and does not depend on the number of requests sent to the system per second. However, they show the opposite when analyzing the response delay. Specifically, we recorded anomalies in two specific cases (i.e., initialization and query), while the latency of on-chain data updates is quite balanced and does not depend

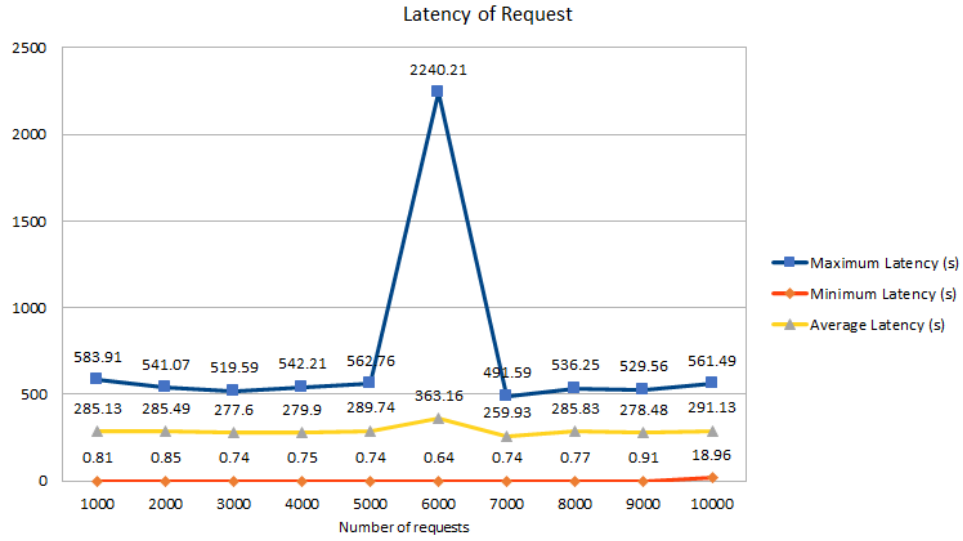


Fig. 8. The Latency of the Data Initialization Functions.

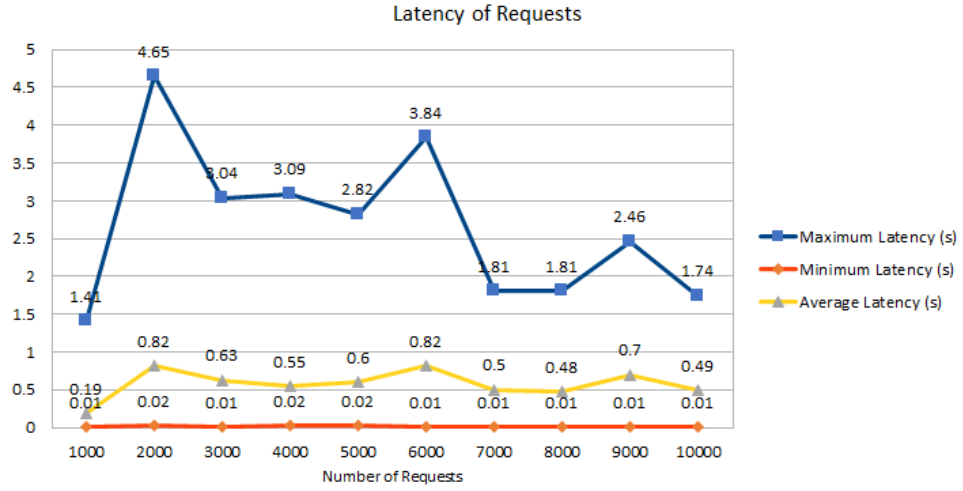


Fig. 9. The Latency of the Request Data Functions.

on the number of responses per second. Specifically, in the case of 6000 requests/second, the latency spikes and drops immediately for the next case. We assume that a problem may occur in the in-chain processing, severely affecting the system's response time. Proof of our claim is that latency tends to be stable from the 7000 requests/second (see Fig. 8). To prove the above statement, we closely observed the peer pairs with transactions during operation and discovered a serious error affecting the whole system. Specifically, when a peer belonging to a specific transaction has not completed the execution request, the whole system must wait until that peer completes, seriously affecting the whole system. We further assume that executing requests on a system that is limited to our hardware equipment is responsible for the above risk.

To answer this question, we continue to observe in the query scenario (i.e., Fig. 9) and update (i.e., Fig. 10) on the data. whether on-chain. Our assumption is not entirely correct for two reasons. The first reason comes from a detailed assessment of client data queries' response time. Specifically, the 2000 milestone is unusually higher than the other observed milestones. This leads to concerns about process interruptions and system-wide effects (since 2000 requests/seconds is not a theoretically alarming milestone). However, when we observed two other anomalies in the same scenario (i.e., 6000 and 9000 requests Fig. 9), we noticed a gradual decrease in latency. This proves the opposite of the assumption above that latency is inversely proportional to the system's processing speed (i.e. the more processing, the more latency increases and vice versa). This proves that the latency of the whole system does not depend on the system configuration but is affected by the sequence processing (i.e., priority).

In the third scenario, we ignore the peers when there is a

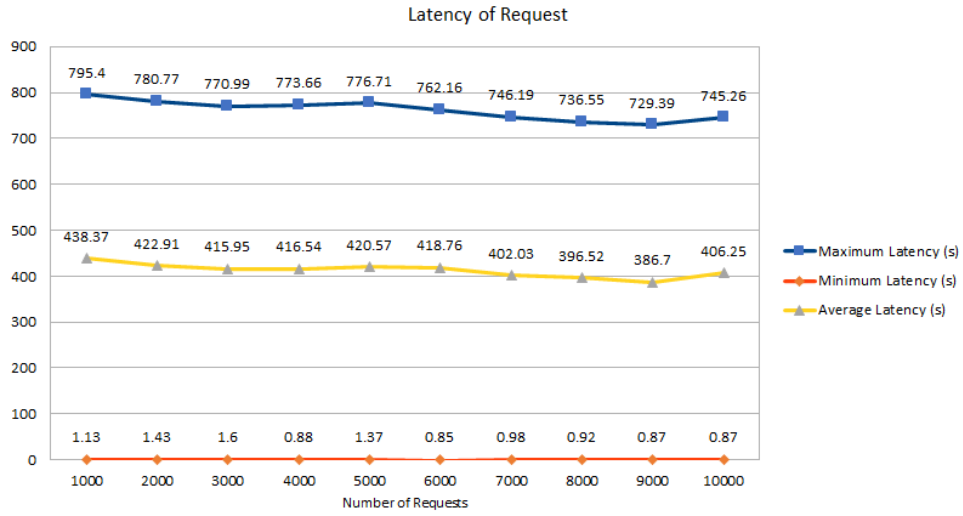


Fig. 10. The Latency of the Update Data Functions.

problem in a corresponding transaction to substantiate these remarks. Fig. 10 proves that our assessment is correct as changes in the number of requests do not affect system latency.

B. Security and Privacy Discussion

The blockchain has a key role in the modern system, which satisfies the transparency and availability conditions for the supply chain [34], [35], [14]. However, the blockchain-based system has some troubles documented in [36], particularly in the security and privacy perspective. To enhance the security and privacy matters in the blockchain-based blood donation system, we consider the authorization for the parties within the same transaction and the flexibility in the clinical environment.

For the authorization factor, we will take the advantage of attribute-based access control (ABAC) [37], [38] to handle the access control process. The primary advantage of this model is that only accepted users can get to the released data. Moreover, the query rewriting can be utilised for a complicated context where the released data is disseminated to numerous users [39], [40]. Eventually, a few methodologies split the original policy into sub-policy [41], [42] (i.e., public and private policy) to assure that the information is only accessed through permission even the parties in the same transaction.

Finally, comparing to the previous papers, this is the first approach supplying the proof-of-idea to target blood donation management. In other words, several papers pay attention to the general problems in the healthcare system instead of focusing on the specific concern, i.e., blood management. This paper underlined BloodChain as a technology to manage the benefits of this system. We also show the outcome of our suggested model derived from some experiments.

VII. Conclusion

The article applies the benefits of Blockchain technology (i.e., transparency, decentralized storage) to propose a blood and blood product processing process based on the limitations of the current traditional process in Vietnam. Vietnam. The paper provides a proof-of-concepts based on the Hyperledger Fabric platform, which stores information about blood and its products during the storage and transport processes. The information is stored transparently for easy verification in transit and storage. This is an initial effort in applying the benefits of blockchain technology in designing and managing the supply chain of blood and its products for Vietnam in particular and developing countries in general.

In future work, we aim to manage stakeholders based on constraints defined in the form of Smart Contracts. Moreover, this research result is only the first step to build a system based on blockchain technology in a real environment. Therefore, we aim to deploy the proposed model for exporting in more complex scenarios where there are multiple-role of users and off-chain executions (i.e., out of scope for current version) processes of the medical facilities.

References

[1] H. T. Le, K. L. Quoc, T. A. Nguyen, K. T. Dang, H. K. Vo, H. H. Luong, H. Le Van, K. H. Gia, L. V. Cao Phu, D. Nguyen Truong Quoc *et al.*, "Medical-waste chain: A medical waste collection, classification and treatment management by blockchain technology," *Computers*, vol. 11, no. 7, p. 113, 2022.

[2] B. Colvin, J. Astermark, K. Fischer, A. Gringeri, R. Lassila, W. Schramm, A. Thomas, J. Ingerslev, and I. D. W. Group, "European principles of haemophilia care," *Haemophilia*, vol. 14, no. 2, pp. 361–374, 2008.

[3] P. Sullivan, "Developing an administrative plan for transfusion medicine—a global perspective," *Transfusion*, vol. 45, pp. 224S–240S, 2005.

[4] J. Chapman, "Unlocking the essentials of effective blood inventory management," *Transfusion*, vol. 47, pp. 190S–196S, 2007.

[5] "Importance of the blood supply," https://www.redcrossblood.org/donate-blood/how-to-donate/how-blood-donations-help/blood-needs-blood-supply.html, accessed: 2022-04-30.

[6] E. Lownik, E. Riley, T. Konstenius, W. Riley, and J. McCullough, "Knowledge, attitudes and practices surveys of blood donation in developing countries," *Vox sanguinis*, vol. 103, no. 1, pp. 64–74, 2012.

[7] J. J. Mammen, E. S. Asirvatham, J. Lakshmanan, C. J. Sarman, A. Pandey, V. Ranjan, B. Charles, T. Mani, S. D. Khaparde, S. Upadhyaya *et al.*, "The clinical demand and supply of blood in india: A

national level estimation study," *Plos one*, vol. 17, no. 4, p. e0265951, 2022.

[8] "Blood donation frequently asked questions," https://www.mayoclinic.org/blood-donor-program/faq, accessed: 2022-04-30.

[9] H. X. Son, M. H. Nguyen, N. N. Phien, H. T. Le, Q. N. Nguyen, V. Dinh, P. Tru, and P. Nguyen, "Towards a mechanism for protecting seller's interest of cash on delivery by using smart contract in hyperledger," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 4, pp. 45–50, 2019.

[10] N. Duong-Trung, X. S. Ha, T. T. Phan, P. N. Trieu, Q. N. Nguyen, D. Pham, T. T. Huynh, and H. T. Le, "Multi-sessions mechanism for decentralized cash on delivery system," *Int. J. Adv. Comput. Sci. Appl*, vol. 10, no. 9, 2019.

[11] H. T. Le, N. T. T. Le, N. N. Phien, and N. Duong-Trung, "Introducing multi shippers mechanism for decentralized cash on delivery system," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, 2019.

[12] N. T. T. Le, Q. N. Nguyen, N. N. Phien, N. Duong-Trung, T. T. Huynh, T. P. Nguyen, and H. X. Son, "Assuring non-fraudulent transactions in cash on delivery by introducing double smart contracts," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 677–684, 2019.

[13] H. T. Le, L. N. T. Thanh, H. K. Vo, H. H. Luong, K. N. H. Tuan, T. D. Anh, K. H. N. Vuong, H. X. Son *et al.*, "Patient-chain: Patient-centered healthcare system a blockchain-based technology in dealing with emergencies," in *International Conference on Parallel and Distributed Computing: Applications and Technologies*. Springer, 2022, pp. 576–583.

[14] H. X. Son, T. H. Le, N. T. T. Quynh, H. N. D. Huy, N. Duong-Trung, and H. H. Luong, "Toward a blockchain-based technology in dealing with emergencies in patient-centered healthcare systems," in *International Conference on Mobile, Secure, and Programmable Networking*. Springer, 2020, pp. 44–56.

[15] N. Duong-Trung, H. X. Son, H. T. Le, and T. T. Phan, "Smart care: Integrating blockchain technology into the design of patient-centered healthcare systems," in *Proceedings of the 2020 4th International Conference on Cryptography, Security and Privacy*, 2020, p. 105–109.

[16] ——, "On components of a patient-centered healthcare system using smart contract," in *Proceedings of the 2020 4th International Conference on Cryptography, Security and Privacy*, 2020, p. 31–35.

[17] H. T. Le, T. T. L. Nguyen, T. A. Nguyen, X. S. Ha, and N. Duong-Trung, "Bloodchain: A blood donation network managed by blockchain technologies," *Network*, vol. 2, no. 1, pp. 21–35, 2022.

[18] N. T. T. Quynh, H. X. Son, T. H. Le, H. N. D. Huy, K. H. Vo, H. H. Luong, K. N. H. Tuan, T. D. Anh, N. Duong-Trung *et al.*, "Toward a design of blood donation management by blockchain technologies," in *International Conference on Computational Science and Its Applications*. Springer, 2021, pp. 78–90.

[19] M. S. Shahbaz, R. Z. RM, M. F. Bin, and F. Rehman, "What is supply chain risk management? a review," *Advanced Science Letters*, vol. 23, no. 9, pp. 9233–9238, 2017.

[20] O. Lavastre, A. Gunasekaran, and A. Spalanzani, "Effect of firm characteristics, supplier relationships and techniques used on supply chain risk management (scrm): an empirical investigation on french industrial firms," *International Journal of Production Research*, vol. 52, no. 11, pp. 3381–3403, 2014.

[21] A. Nagurney, A. H. Masoumi, and M. Yu, "Supply chain network operations management of a blood banking system with cost and risk minimization," *Computational management science*, vol. 9, no. 2, pp. 205–231, 2012.

[22] N. Armaghan and N. Pazani, "A model for designing a blood supply chain network to earthquake disasters (case study: Tehran city)," *International Journal for Quality Research*, vol. 13, no. 3, pp. 605–624, 2019.

[23] M. Eskandari-Khanghahi, R. Tavakkoli-Moghaddam, A. A. Taleizadeh, and S. H. Amin, "Designing and optimizing a sustainable supply chain network for a blood platelet bank under uncertainty," *Engineering Applications of Artificial Intelligence*, vol. 71, pp. 236–250, 2018.

[24] D. Delen, M. Erraguntla, R. J. Mayer, and C.-N. Wu, "Better management of blood supply-chain with gis-based analytics," *Annals of Operations Research*, vol. 185, no. 1, pp. 181–193, 2011.

[25] H. H. Luong, T. D. Anh, K. N. H. Tuan, and H. X. Son, "Ioht-mba: An internet of healthcare things (ioht) platform based on microservice and brokerless architecture," 2021.

[26] L. N. T. Thanh, N. N. Phien, H. K. Vo, H. H. Luong, T. D. Anh, K. N. H. Tuan, H. X. Son *et al.*, "Sip-mba: A secure iot platform with brokerless and micro-service architecture," 2021.

[27] N. T. T. Lam, H. X. Son, T. H. Le, T. A. Nguyen, H. K. Vo, H. H. Luong, T. D. Anh, K. N. H. Tuan, and H. V. K. Nguyen, "Bmdd: A novel approach for iot platform (broker-less and microservice architecture, decentralized identity, and dynamic transmission messages)," *International Journal of Advanced Computer Science and Applications*, 2022.

[28] F. Alharbi, "Progression towards an e-management centralized blood donation system in saudi arabia," in *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*. IEEE, 2020, pp. 1–5.

[29] S. Lakshminarayanan, P. Kumar, and N. Dhanya, "Implementation of blockchain-based blood donation framework," in *International Conference on Computational Intelligence in Data Science*. Springer, 2020, pp. 276–290.

[30] K. Toyoda, P. T. Mathiopoulos, I. Sasase, and T. Ohtsuki, "A novel blockchain-based product ownership management system (poms) for anti-counterfeits in the post supply chain," *IEEE access*, vol. 5, pp. 17 465–17 477, 2017.

[31] X. S. Ha, T. H. Le, T. T. Phan, H. H. D. Nguyen, H. K. Vo, and N. Duong-Trung, "Scrutinizing trust and transparency in cash on delivery systems," in *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*. Springer, 2020, pp. 214–227.

[32] X. S. Ha, H. T. Le, N. Metoui, and N. Duong-Trung, "Dem-cod: Novel access-control-based cash on delivery mechanism for decentralized marketplace," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2020, pp. 71–78.

[33] K. L. Quoc, H. K. Vo, L. H. Huong, K. H. Gia, K. T. Dang, H. L. Van, N. H. Huu, T. N. Huyen, L. Van Cao Phu, D. N. T. Quoc *et al.*, "Sssb: An approach to insurance for cross-border exchange by using smart contracts," in *International Conference on Mobile Web and Intelligent Information Systems*. Springer, 2022, pp. 179–192.

[34] L. Campanile, M. Iacono, A. H. Levis, F. Marulli, and M. Mastroianni, "Privacy regulations, smart roads, blockchain, and liability insurance: Putting technologies to work," *IEEE Security & Privacy*, 2020.

[35] N. Kshetri, "Blockchain's roles in strengthening cybersecurity and protecting privacy," *Telecommunications policy*, vol. 41, no. 10, pp. 1027–1038, 2017.

[36] A. P. Joshi, M. Han, and Y. Wang, "A survey on security and privacy issues of blockchain technology," *Mathematical foundations of computing*, vol. 1, no. 2, p. 121, 2018.

[37] N. M. Hoang and H. X. Son, "A dynamic solution for fine-grained policy conflict resolution," in *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, 2019, pp. 116–120.

[38] H. X. Son and N. M. Hoang, "A novel attribute-based access control system for fine-grained privacy protection," in *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, 2019, pp. 76–80.

[39] H. X. Son, T. K. Dang, and F. Massacci, "Rew-smt: a new approach for rewriting xacml request with dynamic big data security policies," in *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*. Springer, 2017, pp. 501–515.

[40] S. H. Xuan, L. K. Tran, T. K. Dang, and Y. N. Pham, "Rew-xac: an approach to rewriting request for elastic abac enforcement with dynamic policies," in *2016 International Conference on Advanced Computing and Applications (ACOMP)*. IEEE, 2016, pp. 25–31.

[41] H. X. Son and E. Chen, "Towards a fine-grained access control mechanism for privacy protection and policy conflict resolution," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, 2019.

[42] Q. N. T. Thi, T. K. Dang, H. L. Van, and H. X. Son, "Using json to specify privacy preserving-enabled attribute-based access control policies," in *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*. Springer, 2017, pp. 561–570.

Letter-of-Credit Chain: Cross-Border Exchange based on Blockchain and Smart Contracts

Khoi Le Quoc[1], Phuc Nguyen Trong[2], Hieu Le Van[3], Hong Khanh Vo[4], Luong Hoang Huong[5], Khoa Tran Dang[6], Khiem Huynh Gia[7], Loc Van Cao Phu[8], Duy Nguyen Truong Quoc[9], Nguyen Huyen Tran[10], Huynh Trong Nghia[11], Bang Le Khanh[12], Kiet Le Tuan[13]
FPT University, Can Tho City, Viet Nam

***Abstract*— The exchange of goods between countries is growing, contributing to the promotion of logistics-related technologies. More and more systems are adopting advances in science and engineering to reduce manual handling steps, thereby reducing transit time. Letter-of-Credit (LOC) is a standard method where the parties involved will enter into agreements for the sale and exchange of goods. Specifically, each party will receive a set of original documents and does not need to meet face-to-face under the bank's witness. The process brings many benefits in terms of time and reduces records processing. However, the system faces a lot of risks when one of the parties is dishonest. On the other hand, the traditional LOC systems face a lot of risks related to the transparency of information about the goods, and also the supplier may lose the goods (e.g., 4/100 Vietnamese cashew nut containers are lost. stuck in Italy) or deposits in the hands of shipping companies (e.g., GNN Express - Vietnam) and many more. To this end, many research directions have exploited blockchain technology and smart contracts. Specifically, all information related to the transaction between the supplier and the demander including package, time, and delivery location. However, there needs to be a mechanism to ensure the smooth implementation of smart contracts, specifically for sanctioning when there is a conflict between a supplier and a demander. This role should be considered for the transaction manager, who directly designs and is responsible for their smart contracts. Currently, there is no mechanism to guarantee all interests of the parties involved in non-bank transactions. To increase the processing capacity and integrate with the Blockchain system, we propose the Letter-of-credit Chain that defines the agreements between the parties in international trade. We also deploy the proof-of-concept of the Letter-of-credit Chain on the three EVM-supported platforms (i.e., under ERC20), namely, Ethereum, Binance Smart Chain, and Fantom. By evaluating the actual execution of Gas for each platform, we found that our proposed model had the cheapest fee when deployed on the Fantom platform. Finally, we share the deployment/implementation of these platforms' proof-of-concept to encourage further future research.**

***Keywords*—Letter-of-Credit; blockchain; smart contract; authorization; Ethereum; Fantom; Binance smart chain**

I. Introduction

It is undeniable that the development of technology has changed almost completely the approach to a business organization as well as a management strategy to meet customer requirements [1]. The exchange of goods no longer takes place in a narrow area limited in area, but instead, companies can transport and export/import goods from all over the world. Traditional international trade models were originally built on trust. Specifically, in the first stage, the supplier will bring the goods to another city or country to find a demander. It is clear that this model is very risky for both the supplier and the demander. In particular, demanders may purchase inferior products because all constraints are not controlled; on the supplier's part, they may lose the entire goods if the demander refuses to pay or the product is past its expiry date due to the time-consuming shipping process [2]. In order to solve the risks of transporting the goods, the supplier will authorize the transport company when transporting the goods to the demander (i.e., minimizing risks in transportation and transit time), i.e. that the parties can communicate indirectly through intermediaries instead of having to meet face-to-face - this also reduces the costs incurred. Due to the growing and expanding trade, where the need to transport goods increases and the transit time decreases. Therefore, both parties will authorize a trusted third party called a transaction manager (e.g. a bank). All contractual requirements and constraints must be accepted by all three parties (i.e. supplier, demander and third party). The role of the intermediary is assigned to the Bank (called Letter-of-Credit - LOC). Specifically, the demander is provided with an economic guarantee from the bank that grants credit to the exporter of the goods [3]. The supplier receives the money only if and only when the demander receives the goods and provides all statements related to the shipment confirmed by the transaction manager. An important disadvantage of the traditional LOC model is that it is easy for suppliers to lose goods if they work with the untrusted transaction manager and malicious demanders. One of the most examples of this problem is introduced in Section IV, which present the cashew nut export from Vietnam to Italy in 2022[1].

E-commerce has enhanced the process of transporting goods across borders thanks to the exchange of routes through e-commerce platforms. This process is made faster and brings many benefits to both the supplier and the demander. A series of e-commerce platforms (e.g., Amazon, Alibaba) have largely changed users' shopping habits, while shipping companies (e.g., FedEx, ASL) have also accelerated the conversion process. goods. The freight conversion process is based on a Cash-on-Delivery (COD) shipping company, where the carrier will play an extremely important role in delivering and receiving the demander's funds. Payment to the supplier is the responsibility of the carrier. Most shipping companies will keep the demander's money before handing it over to the supplier. If the shipping company goes bankrupt or refuses to pay, the supplier will lose money [4]. In fact, a series of shipping companies have appropriated the supplier's money (e.g., GNN Express).

[1]https://english.vov.vn/en/economy/vietnam-requests-italy-to-investigate-suspected\ \-cashew-nut-export-scam-post931226.vov

To solve these challenges, several Blockchain-based approaches have been developed to replace the traditional LOC model. These new systems provide a secure platform for both suppliers and demanders, aka decentralized marketplaces/exchanges [5]. These protocols focus on dealing with issues related to suppliers, demanders, and carriers (see details in the III section). However, these models still do not meet the requirements of international trade, where the role of the transaction manager is extremely important [6]. To solve this problem, we introduce Letter-of-Credit Chain, a system based on Blockchain and smart contracts to solve the insurance problem for cross-border exchange. Letter-of-Credit Chain builds three main user groups, including suppliers, demanders and transaction manager. This model consists of eight main steps, from the supplier creating the package/goods to the order being delivered to the demander and the order arrival. Besides, we develop a role-based access control model (i.e., authorization) to define logical constraints in smart contracts to maintain the stable operation of the system. To define the logical binding between stakeholders on smart contracts, we also exploit the Solidity. To evaluate the Chain of Credit, we deployed a test model for our proof-of-concept on all three of the most popular platforms that currently support the Ethereum Virtual Machine (EVM) environments, including Ethereum, BNB Smart Chain, and Fantom (see [7] for further info). To support the current letter-of-credit system (i.e., International Trade), we share our proof-of-concept implemented on all three platforms.

Following this section, we present the background of blockchain technology and its platform in Section II. Whereas, Section III describes the summary, limitations and challenges of the current approaches. Then, the two next sections define the problem statement of the traditional model of the Letter-of-Credit approach and also introduce the architecture of Letter-of-Credit Chain Architecture based on the blockchain and smart contract in Sections IV and V, respectively. Section VI describes the proof-of-concept of the Letter-of-Credit Chain, i.e., data structure, execution algorithm, and authorization. Last but not least, Section VII focuses on the effectiveness proof via the evaluation process based on deploying Letter-of-Credit Chain in the ETH, BNB, and Fantom platforms. Finally, we continue with the conclusion and the future work of the article in the last section.

II. Background

A. Blockchain Technology

Blockchain was popular after the introduction of Bitcoin by Nakamoto in 2008 [8] and is usually represented as a transparent, trusted, and decentralized ledger. The blockchain-based system manages transaction data on multiple computers simultaneously on a peer-to-peer network. Therefore, it creates a secure connection between the transacting parties (i.e., receiver and money transmitter) without the need for a traditional third party (e.g., a bank) [9].

The most popular types of blockchains today include Public, Private, and Hybrid (Called Consortium). In the first type, the two best-known examples of public blockchains currently are Bitcoin and Ethereum. Any users (including hidden ones) could join to Blockchain network to view content, execute a new transaction, or check the integrity of the existing blocks. For private blockchains, some common examples of this type include GemOS, MultiChain, Ripple, and Eris. Unlike the Public type, they only support authorized users who can join the network as well as execute, check transactions to the block or create a new block [10]. Combining the two, a semi-private (called Consortium) blockchain is defined as the boundary between public and private ones. It strives to achieve outstanding characteristics in each category - specifically, Consortium blockchains are often deployed for enterprises to ensure their security and interact with their partners for better business. Two famous examples of hybrid blockchains are Hyperledger Fabric [11] and Ethereum [12] (i.e., which allows the creation of Golang-based federated blockchains).

B. Blockchain Platform

1) Ethereum: Ethereum [13] is a decentralized blockchain platform for running smart contracts with the support of the Solidity programming language. Similar to other high-level languages (e.g., Java), Ethereum is executed by the Ethereum Virtual Machine (EVM). Ethereum supports decentralized finance (DeFi) protocols where smart contract-based constraints are provided.

2) Hyperledger Fabric: Hyperledger Fabric [11] is open-source enterprise-grade permission designed for large-scale commerce. This platform is designed based on distributed hyper-ledger mechanism and supports both public and private blockchain platforms simultaneously. Hyperledger Fabric and Ethereum together perfect Turing. However, instead of executing smart contracts on the EVM virtual machine like Ethereum, Hyperledger code is executed in Docker containers called ChainCode. It allows developer applications to deploy smart contracts with minimal overhead. Another advantage over Ethereum is that it supports high-level programming languages (i.e., Java and Go) instead of relying on Solidity. With this advantage, Fabric has facilitated the development and maintenance of the platform. By not having to switch to a new language, Fabric has helped to reduce operating costs (e.g., system maintenance, information storage and querying within the blockchain).

C. Smart Contracts

A smart contract (Ethereum) or chaincode (Hyperledger Fabric) is a term that describes a set of protocols that assist developers in defining terms and agreements in transactions between the parties to the contract. The entire process of smart contracts is not dependent on external interference and is performed automatically based on the support of Blockchain technology. In a Dapp, the terms and constraints of a smart contract are recorded in the language of a computer and are equivalent to a legal contract.

1) Characteristics: The Smart Contract routine has the following characteristics:

- **Distributed**: Replicated and distributed in all nodes of the Ethereum network. This is one difference from other solutions based on centralized servers.
- **Deterministic**: Only take actions that they are designed to perform if the conditions are satisfied. Be-

sides, the results of Smart Contracts remain the same no matter who the executor is.

- **Automate**: Able to automate all kinds of tasks, and it works like a self-executing program. However, in most cases, if the Smart Contract is not activated, it will remain "inactive" and will not perform any action.
- **Non-modifiable**: Smart Contract cannot be modified after deployment. They can only be "deleted" if this function has been added before. Therefore, it can be said that Smart Contract is like an anti-forgery code.
- **Customizable**: Before deployment, Smart Contracts can be encoded in different ways. So, they can be used to create many types of decentralized applications (Dapps). Ethereum is a blockchain that can be used to solve any computational problem (Turing complete).
- **No need to rely on trust**: Two or more parties in a contract can interact through a Smart Contract without knowing or trusting each other. In addition, blockchain technology ensures the accuracy of data.
- **Transparency**: Since Smart Contracts are based on a public blockchain, no one can change their source code, although anyone can view it.

2) How Smart Contracts Work: The working principle of a smart contract can be compared to a vending machine. It only automatically executes pre-programmed commands.

First, assets and contract terms are both encrypted and transferred into a block on the Blockchain. Then this smart contract will be distributed and copied by the nodes working on that platform. After receiving the deployment order, the contract will be deployed according to the predetermined terms. Simultaneously, the smart contract will also automatically check the implementation of the commitments stated in the agreement.

- Cost savings: Pay a minimal fee to the blockchain network, saving fees.
- Flexibility: The terms in the contract are handled flexibly and efficiently for the user.
- Transparency, clarity: all payment transactions can be traced, but payment transactions will not be reversed at all, and all transactions will be recorded on the blockchain with extreme clarity.
- High Reliability: Once the contract is completed, no one or a party can interfere in the execution and negotiation of the contract.
- Fast, convenient: can set up and execute a contract in seconds, install for many people simultaneously, and use it many times.

3) Solidity: Solidity is an object-oriented, high-level language for implementing smart contracts. Smart contracts are programs that oversee the conduct of records inside the Ethereum state. Solidity was influenced by JavaScript, C++, and Python and is intended to focus on the Ethereum Virtual Machine (EVM).

Solidity is statically composed and upholds inheritance, libraries, and complex client-characterized types, among different highlights. With Solidity, you can make contracts for utilizations like democratic, crowdfunding, dazzle barters, and multi-signature wallets.

4) Web3.js: Main steps in creating blockchain applications with Ethereum:

- Innovative contract development - composing code that gets sent to the blockchain with the Solidity programming language.
- It is creating sites or customers that cooperate with the blockchain - composing code that peruses and contains information from the blockchain with smart contracts.

Web3.js empowers you to satisfy the following duty: creating customers that communicate with The Ethereum Blockchain. An assortment of libraries permits you to perform activities like sending Ether starting with one record and then onto the next, peruse and compose information from shrewd agreements, make smart contracts, and thus significantly more.

If you have a web advancement foundation, you may have utilized jQuery to settle on Ajax decisions to a web worker. That is a decent beginning stage for understanding the capacity of Web3.js. Rather than using jQuery to peruse and compose information from a web worker, you can utilize Web3.js to peruse and keep in touch with The Ethereum Blockchain.

Web3.js converses with The Ethereum Blockchain with JSON RPC, which means "Remote Procedure Call" convention. Ethereum is a distributed organization of hubs that stores a duplicate of all the information and code on the blockchain. Web3.js permits us to make solicitations to an individual Ethereum hub with JSON RPC to peruse and compose information for the organization. It's similar to utilizing jQuery with a JSON API to peruse and manage data with a web worker.

5) Remix: The remix is a Solidity IDE used for writing, compiling, and debugging Solidity code. Solidity is a high-level, contract-orientated programming language for writing clever contracts. It is affected by popular languages such as C++, Python, and JavaScript.

Ethereum is a general-purpose blockchain, which is more suitable for using advanced scripts (also known as smart contracts) to describe business logic. Ethereum is developing as a decentralized or global computer that combines blockchain functions with a broader perspective. As a reliable machine with a complete Turing contract engine.

Benefits of using remixed IDE to compile and deploy smart contracts:

- Compile the contract in Remix IDE.
- See a few warnings issued with the help of using a compiler while high-quality practice is not followed.
- Contract implementation on JavaScript EVM (Ethereum Virtual Machine).
- Make transactions with the implemented contracts.
- See example reading and writing in the IDE terminal.

D. Our Selection Platform

As mentioned in the Introduction section, this article deploys the proof-of-concept of the Letter-of-credit Chain on three platforms, namely, Ethereum, Binance Smart Chain, and Fantom[2].

1) Binance Smart Chain: Binance Smart Chain (BSC) is an enhanced version of the original Binance Chain version. BSC is designed to be a parallel platform to the first version. Similar to Ethereum, BSC offers Dapp developers options to support smart contracts and can be deployed on other Blockchain platforms that support EVM.

2) How does Binance Smart Chain Work?: BSC applies a hybrid model of Proof of Authority and Proof of Stake - called Proof of Staked Authority (PoSA). Validators for the BNB system will put a certain amount of BNB into the system and receive a bonus after each successful validation.

Binance Chain and Binance Smart Chain are cross-chain compatible and designed to be completely in sync. With BSC, assets can be moved between blocks thanks to the fast transaction capabilities of the original version and smart contracts of the improved version (i.e., EVM integration). Specifically, Binance Chain supports two tokens (BEP-2 and BEP-8). In addition, Binance Chain can also be swapped with Smart Chain BEP-20 tickets. So, thanks to ERC-20 contract compatibility, DApp developers on other EVM-enabled platforms can switch to Binance Smart Chain relatively quickly.

III. Related Work

Several protocols exploited the blockchain system's advantage to improve their transaction among the peers or components. Some of them considered Cash-on-Delivery (COD) model [14], [15] (i.e., suppliers, demanders), the medical care system [16], [17], [18] (i.e., doctors and patients), health care emergency situation [19], [20] (i.e., medical staff and the patients' friends or relatives), or blood donation - humanitarian blood transfusion (donors and recipients) [21], [22] and much more.

To prove the improvement of Blockchain for traditional shipping (i.e., Letter-of-Credit shipping (called LOC) or COD) Ha et al. [6] described the current shipping system faced the massive drawback (e.g., dependence on trusted third parties, goods/order management, complicated payment processes among the parties in the same peer or ecosystem, losing the package and deposit for supplier and demander, respectively. To this end, Le et al. [23] suggested that blockchain technology could fill these gaps via the usage of smart contracts and decentralized management in COD general and LOC special. For instance, the Ethereum ecosystem proposed a method called `localEthereum` which is introduced to support the transaction or DeFi Dapp between the suppliers and demanders [24]. Similarly, `OpenBazaar` [25] developed based on the `localEthereum` extension, in which this protocol defined the demander and supplier-sponsored. However, compared to `localEthereum`, the main difference between `OpenBazaar` was that the `OpenBazaar` involved the three parties: the supplier, the demander, and the moderator (i.e., a new role in control).

Moreover, a new protocol targeted at helping transport products from suppliers to demanders [26] exploited the ETH-based transaction to propose a COD/LOC mechanism. In particular, this tool considers the new actor (i.e., shipper) rather than focusing on only the transaction between the demander and supplier as in the previous approaches above. However, this approach still has the main drawback is that it required trusted behavior from the shipper (i.e., new role) not only in this task but also the interaction between the supplier and demander (i.e., how can be proof of this level is still the open question). Furthermore, there are impossible to assume that the stakeholders are trusted for all their behavior in the system/network. Hence, this is the main limitation of these approaches.

To increase the shipper's role in the blockchain-based approaches, some studies (e.g., [27] and [28]) introduced carriers in the decentralized marketplace, rather than focusing only on demanders and suppliers. These works developed the transportation processes as well as provided the mechanisms to promote and ensure the benefit of the stakeholders. These models also penalized any parties who intentionally commit fraud; therefore, the demanders' and suppliers' interests are enhanced. However, the scope of these models' application has limited to the distance of shipping where the transaction occurs in the same city or at most in the same country. Last but not least, they have not considered the conflict issues among the parties in the same transaction, for instance, suppliers and demanders.

Compared to the existing works, our proposed model (i.e., Letter-of-Credit Chain) introduces a trusted international trade channel that connects all the suppliers and demanders around the World (among the different countries). We aim to reduce the conflict among the parties in the system to solve the two case studies, namely, GNN Express as well as cashew nut export problems as present in the Introduction section.

IV. Problem Statement

Fig. 1 describes the problem of the cashew nut export from Vietnam to Italy in 2022.

The Vietnamese enterprise (Exporter) first signs a contract with the Italian importer to specify the contract's bank and port of destination. The Exporters then deliver goods to shipping lines and receive "original documents". Apter that, the Exporter receive the original documents and bring them to a bank in Vietnam (a.k.a Exporter's bank) to ask this bank to collect money. The Exporter's bank next send this set of original documents to the importer's bank in Italy. At this point, the Italian importer's bank will pay the exporter's bank and give this original document to the Italian importer for them to receive the goods.

However, the problem with this is that the importer's bank in Italy did not receive the original set of documents, but they get photocopies instead. Thus, this bank refuses to transfer money.

[2]Since selected platforms supporting the EVM are similar in the execution process (BNB Smart Chain and Fantom), this section gives summarize of the Binance Smart Chain in a nutshell.

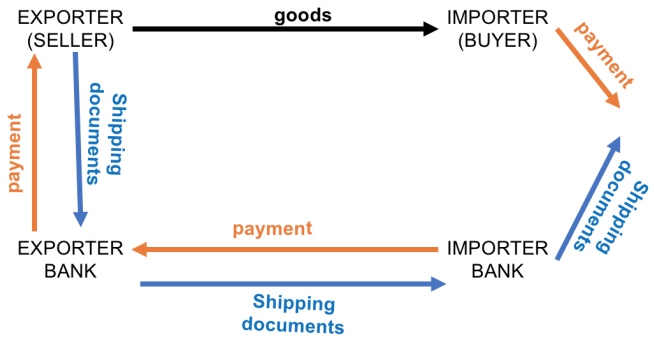


Fig. 1. The Problem Statement of the Cashew Nut from Vietnam to Italy in 2022.

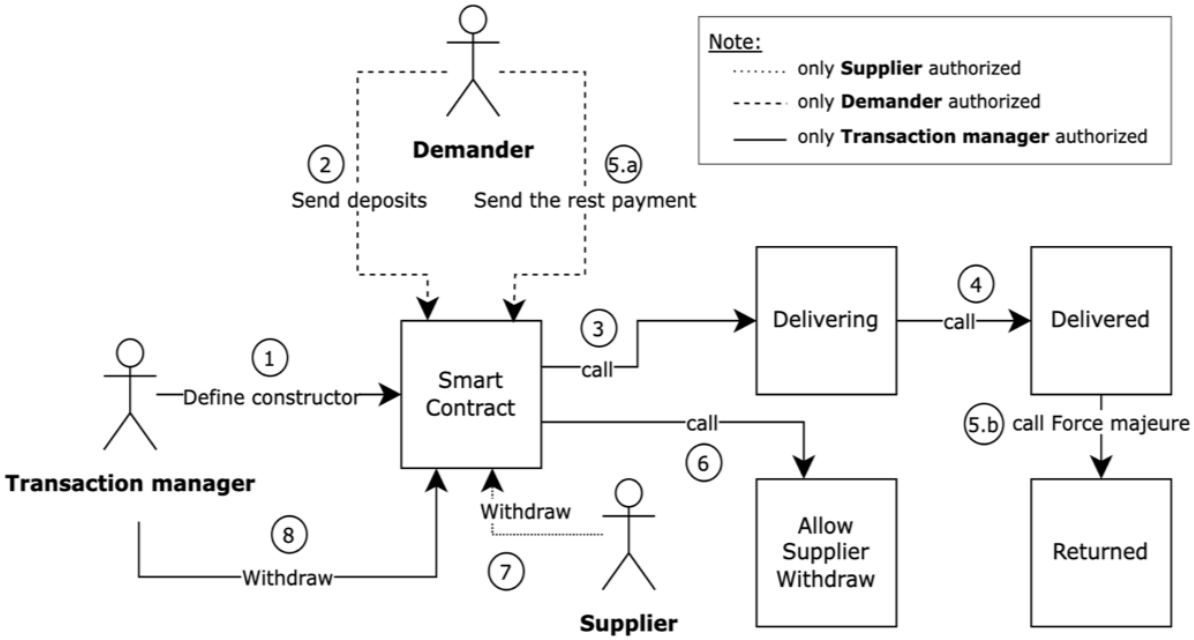


Fig. 2. Letter-of-Credit Chain Architecture.

V. Letter-of-Credit Chain Architecture

Fig. 2[3] shows Letter-of-Credit Chain architecture including three main components, namely supplier, demander, and transaction manager.

- *supplier*: The party who provides goods or services to another organization. In Letter-of-Credit Chain architecture, the Supplier can withdraw the money after the goods are delivered to Demander via the smart contract.
- *demander*: Who makes requests for an item or service in the platform economy. In this architecture, the Demander must send the amount of deposit as well as the remaining money to the Suppier via smart contract. In the case that Demander won't send the remaining amount, their deposit is lost. We also consider this point in our experiments (see Section VII).
- *transaction manager or Trustee*: the third party who manages both traditional contracts and smart contracts. *TransactionManager* controls the two flows i.e., the order to Demander and the money to Supplier.

The TransactionManager define the smart contract (a.k.a. the tradition contract) and upload this on the Letter-of-Credit Chain [Step 1]. This can identity the address on the corresponding blockchain network. The second step show that the must be send the deposit to certain smart contract defined by the TransactionManager. In our approach rather than one-time-transfer the Demander can split this as multiple amount; however the sum of these amounts must be higher than 50% of the order's price. The order status is changed to Delivering in the third step, whenever the *TransactionManager* confirm that the shipping function is called. In the next step (i.e., the order delivered) we deploy this as off-chain. In the current version, we do not focus on the off-chain tasks in the whole transportation process. Step 4 calls the delivered function if and only if the order has arrived. We have two cases in this point,

[3]In this model architecture, we do not refer to off-chain tasks.

i.e., the Demander transfer the remaining amount in Step 5(a) or does not transfer in Step 5(b). Regarding to the behavior of the Demander we have two corresponding solutions that are i) if they transfer the remaining, the Supplier can withdraw their money for the goods/order via Allow Supplier Withdraw function in Step 6; otherwise the Demander lost their deposit and this amount automatically send to Supplier as shown in Step 7. Finally, the *TransactionManager* withdraw their amount based on the smart contract fee execution in Step 8 and does not depend on the Demander decision.

VI. Implementation

A. Data Structure

Fig. 3 describes the Letter-of-Credit Chain framework' data structure. We only consider the key information (i.e., Order or Goods) in this paper. The remaining ones are described in our code. Please follow the deployment of our implementation in the three platforms (see Section VII for more detail).

B. Algorithm

Letter-of-Credit Chain framework (i.e., Algorithm 1) executes from top to bottom. First, the TransactionManager add the smart contract code to the blockchain network. This point is created by the TransactionManager, but we can easily to detect the meaning of their requirement. In the next step we set up current_State; State.Created; and balance_Received = 0. When the Supplier uploads their orders to the network and finds the corresponding Demander (see lines 4 to 15) the Demander must sends deposit money to the smart contract. At this point, Algorithm 1 updates balance_Received as well as the smart contract also logs the history of the transaction into payment_Histories parameter. If the order on the delivering (current_State = State.Hold or State.Complete_Payment), the state of current order current_State equal to State.Delivering value (see the if command lines 16 - 19). For the next If condition, the order has delivered, we update current_State = State.Delivered (lines 19-21)

The while command to verify whether the Demander send the remaining or not (see V for more details). Lines 30 and 31 show the Demander paying the remaining amount of money and our process; otherwise, please follow lines 32 and 33. Finally, the Supplier receives their amount (see lines 35-37) as well as the TransactionManager (see lines 38 - 40)

Moreover, the Letter-of-Credit Chain also provides the RBAC service for the three actors in the system. In this service, we allow the authorized partner can call the corresponding function/method. The list of functions is presented in our paper public in [7]. However, the data and meta data of the transaction is still public for all the stakeholders. Please follow our analysis w.r.t security and privacy which describes in Section VII-D

Algorithm 1 Letter-of-Credit Chain Execution

```
1: Input: contract_Name, Transaction_Manager, Demander, Supplier, order_Amount, tax_Amount, deposit_Amount
2: current_State = State.Manager_Withdrawm
3: Begin: set balance_Received = 0; set current_State = State.Created
4: while Demander transfers deposit>=deposit_Amount do do
5:    update balance_Received
6:    storing payment transaction to payment_Histories
7:    if balance_Received<deposit_Amount then
8:       update current_State = State.DEPOSIT
9:    else if balance_Received>=deposit_Amount && balance_Received<order_Amount then
10:      update current_State = State.HOLD
11:   else
12:      update current_State = State.COMPLETE_PAYMENT
13:   end if
14:   storing order's status to orderStatus
15: end while
16: if current_State == State.HOLD or current_State == State.COMPLETE_PAYMENT then
17:   manual update current_State = State.SHIPPING
18: end if
19: if current_State == State.SHIPPING then
20:   manual update current_State = State.SHIPPED
21: end if
22: while receiving the rest payment until the deadline do
23:   if balance_Received == order_Amount then
24:      update current_State = State.COMPLETE_PAYMENT
25:      manual update current_State = State.CAN_WITHDRAW
26:   else
27:      manual update current_State = State.NONPAYMENT
28:   end if
29: end while
30: if current_State == State.CAN_WITHDRAW then
31:   set amount_Withdraw = balance_Received - tax_Amount
32: else if current_State == State.NONPAYMENT then
33:   set amount_Withdraw = balance_Received - 2*tax_Amount
34: end if
35: while Supplier withdrew money do
36:   current_State = State.WITHDRAWN
37: end while
38: while TransactionManager withdrew money do
39:   current_State equal to State.MWITHDRAWN
40: end while
```

VII. Evaluation

A. Environmental Setting

The setting of our environment is shown below:

- Blockchain platform: Binance Smart Chain, ETH, Fantom

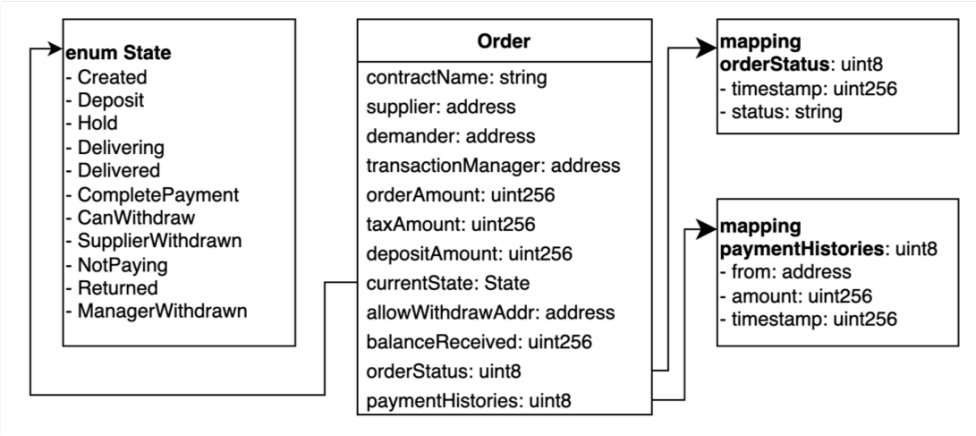


Fig. 3. Data Structure.

- Language: Solidity
- IDE: Remix
- Compiler: 0.8.16+commit.e07a7930a
- Evm version: default
- Gas limit: 3000000
- Optimization: yes
- Open Source License Type: MIT License

B. Results

To prove the effectiveness of Letter-of-Credit Chain, we set up two scenarios (i.e., Supplier pays and not pays the rest of payment) in the three most common EVM Blockchain platforms, namely, Fantom, Ethereum, and Binance Smart Chain. We also provide the source code and installation for the further extension in this topic[4]. We consider the two scenarios (i.e., Demander pay the full payment and Demander does not pay the rest of the payment).

1) Scenario 1: Supplier pays full payment: This scenario will sequentially trigger the states as State.Created, State.Hold, State.Delivering, State.Delivered, State.CompletePayment, State.Can_Withdraw, State.Supplier_Withdrawn, and State.ManagerWithdrawn.[5] In this scenario, the Supplier transferred the payments remaining within the validation time. The result of the first scenario is shown in Table I:

2) Scenario 2: Supplier does not Pay the Rest of Payment: This scenario will trigger the states sequentially as State.Created, State.Hold, State.Delivering, State.Delivered, State.NotPaying, State.Returned, State.Can_Withdraw,State.Supplier_Withdrawn, and State.ManagerWithdrawn. In this scenario, the TransactionManager and the Supplier do not transfer the rest of the payment to the smart contract within the validation time.[6]. The result of the second scenario is shown in Table II.

C. Discussion

Tables I and II present the gas fees for the deployment and execution on the three platforms, i.e., Fantom, Ethereum, and Binance Smart Chain of the Letter-of-Credit Chain[7]. We can easily see that Fantom's smart contract execution fee is the cheapest compared to the other two (i.e., on average 0.08 FTM with $0.02665366). Specifically, the most expensive method is executed with approximately 0.5 FTM ($0.13); whereas the cheapest one is $0.0045 with 0.017 FTM for the two scenarios. On the other hand, ETH is the most expensive, with $8.87 for the most and $0.1 for the least for the two scenarios. Following the ETH is the BSC execution fee of $6.6 to deploy the contract on BSC, which is approximately 17 times higher than Fantom's ones and $0.35 for the cheapest ones for the two scenarios. On average, the gas for all eight functions/methods in the two scenarios is $1.49 and $1.29 for the deployment on Ethereum and Binance Smart Chain, respectively.

D. Security and Privacy Discussion

In this article, we just provide the Letter-of-Credit chain based on the blockchain, which focuses on the decentralize and transparency rather than security and privacy (S&P) issues. In these aspects, we support the basic authorization via the role base access control, in which the right party can be called the corresponding functions. However, the main drawback of the RBAC is that on large-scale systems, RBAC is limited at the #roles. These systems might conflict with or redundancy the new policy; thus, the malicious might attack the system [29]. To this end, we will exploit the attribute-based access control (ABAC) approach, which is introduced by [30] to manage the access control process. In particular, Son et al. [14] define the two-layer of policy for the on-chain and off-chain, respectively. Similarly, some approaches split the original policy into sub-policy (e.g., [31]), i.e., public and private policies to ensure

[4]The implementation/deployment of Letter-of-Credit Chain on:
Fantom platform: https://testnet.ftmscan.com/address/0xF11Fde29e0EB94d977d44c2660F5e0227DC81462#code;
Ethereum platform: https://kovan.etherscan.io/address/0xc3f2e07d850d9131123513e3a106c2ce02b8fa21#writeContract;
Binance Smart Chain platform: https://testnet.bscscan.com/tx/0x236fd512f44fa21148e0f902e72277619e2438d704fe9bfa7d6a8db55f1861b7

[5]see the detail of the function from the previous our publication [7].

[6]see the detail of the function from the previous our publication [7]

[7]redemption value as on 29 August 2022

TABLE I. SCENARIO 1: DEMANDER PAYS FULL PAYMENT

Gas for	Fantom	Ethereum	BNB Chain
Create contract	0.4914263925 FTM ($0.13435057)	0.006000447516801253 ETH ($8.876882043)	0.02384379 BNB ($6.617128601)
Transfer deposits	0.059103118802 FTM ($0.016158143)	0.000422128501969933 ETH ($0.624484242)	0.00266919 BNB ($0.740753609)
Delivering	0.016709903756 FTM ($0.004568304)	0.000079564000556948 ETH ($0.117704596)	0.00078564 BNB ($0.218030813)
Delivered	0.016745106924 FTM ($0.004577928)	0.000079732000558124 ETH ($0.11795313)	0.00078732 BNB ($0.218497046)
Allow Supplier Withdraw	0.027616885296 FTM ($0.007550153)	0.000131498000920486 ETH ($0.194534198)	0.00125798 BNB ($0.34911461)
Supplier Withdraw	0.043235290826 FTM ($0.011820053)	0.000205864001441048 ETH ($0.304549028)	0.00196664 BNB ($0.545781933)
Manager Withdraw	0.027618085404 FTM ($0.007550481)	0.000131503001052024 ETH ($0.194541595)	0.00126903 BNB ($0.352181206)

TABLE II. SCENARIO 2: SUPPLIER DOES NOT PAY THE REST OF PAYMENT

Gas for	Fantom	Ethereum	BNB Chain
Create contract	0.49142639 FTM ($0.13435057)	0.00600045 ETH ($8.87688204)	0.02384379 BNB ($6.61712860)
Transfer deposits	0.05910312 FTM ($0.01615814)	0.00042213 ETH ($0.62448424)	0.00266919 BNB ($0.74075361)
Delivering	0.01670990 FTM ($0.00456830)	0.00007956 ETH ($0.11770460)	0.00078564 BNB ($0.21803081)
Delivered	0.01674511 FTM ($0.00457793)	0.00007973 ETH ($0.11795313)	0.00078732 BNB ($0.21849705)
not Paying	0.02761689 FTM ($0.00755015)	0.00013150 ETH ($0.19453420)	0.00125803 BNB ($0.34912849)
Supplier Withdraw	0.04323529 FTM ($0.01182005)	0.00020586 ETH ($0.30454903)	0.00196664 BNB ($0.54578193)
Manager Withdraw	0.02761809 FTM ($0.00755048)	0.00013150 ETH ($0.19454159)	0.00126903 BNB ($0.35218121)

the data is only accessed via permission even by the parties in the same transaction. Besides, the query rewriting can apply to the complex context where the released data is shared with multiple parties and dynamic context [32], [33]. In particular, the authors proposed the dynamic query VII, which responds the difference value (i.e., details level) based on the user attribute.

VIII. CONCLUSION

This article introduced Letter-of-Credit Chain, which replaces the traditional international trade process based on Letter-of-Credit transactions with edge-cutting ones. Specifically, Letter-of-Credit Chain harnesses the benefits of Blockchain technology and smart contracts to build three user groups: transaction manager, supplier, and demander. Letter-of-Credit Chain's execution process is based on eight steps, from the supplier initiating the order to passing the order on to the demander. Compared to the traditional model of depending on a trusted third party (i.e., a bank), Letter-of-Credit Chain is aimed at decentralized storage, where users can view relevant information regarding their order. If a dispute arises, the role of the transaction manager emerges as the one who decides who should be penalized. Specifically, in the current version, the sanction requirements are all public in the smart contract previously defined by the transaction manager. To demonstrate the effectiveness of the Letter-of-Credit Chain, we implement proof-of-concept on all three popular EVM-enabled platforms today, Fantom, Binance Smart Chain, and Ethereum. The performance comparison section proved that Fantom is a viable platform to deploy for further studies.

Regarding future research directions, we plan to develop the current RBAC model that supports more attributes with the application of ABAC. In addition, the privacy scalability of the Letter-of-Credit Chain is also considered by designing a dynamic request mechanism where the return result depends on the requester's role. On the other hand, Letter-of-Credit Chain does not yet include a shipper role, so a serious consideration to extending the current model is urgently needed.

REFERENCES

[1] S.-Y. Wong and K.-S. Chin, "Organizational innovation management: An organization-wide perspective," *Industrial Management & Data Systems*, 2007.

[2] A. Shleifer and R. W. Vishny, "Liquidation values and debt capacity: A market equilibrium approach," *The journal of finance*, vol. 47, no. 4, pp. 1343–1366, 1992.

[3] J. Dolan, "The law of letters of credit," *THE LAW OF LETTERS OF CREDIT, 4th edition*, pp. 07–36, 2007.

[4] D. Waters, *Supply chain risk management: vulnerability and resilience in logistics*. Kogan Page Publishers, 2011.

[5] H. H. Luong, T. K. N. Huynh, A. T. Dao, and H. T. Nguyen, "An approach for project management system based on blockchain," in *International Conference on Future Data and Security Engineering*. Springer, 2021, pp. 310–326.

[6] X. S. Ha, H. T. Le, N. Metoui, and N. Duong-Trung, "Dem-cod: Novel access-control-based cash on delivery mechanism for decentralized marketplace," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2020, pp. 71–78.

[7] K. L. Quoc, H. K. Vo, L. H. Huong, K. H. Gia, K. T. Dang, H. L. Van, N. H. Huu, T. N. Huyen, L. Van Cao Phu, D. N. T. Quoc *et al.*, "Sssb: An approach to insurance for cross-border exchange by using smart contracts," in *International Conference on Mobile Web and Intelligent Information Systems*. Springer, 2022, pp. 179–192.

[8] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Decentralized Business Review*, p. 21260, 2008.

[9] M. A. Uddin, A. Stranieri, I. Gondal, and V. Balasubramanian, "A survey on the adoption of blockchain in iot: Challenges and solutions," *Blockchain: Research and Applications*, vol. 2, no. 2, p. 100006, 2021.

[10] M. Alharby and A. Van Moorsel, "Blockchain-based smart contracts: A systematic mapping study," *arXiv preprint arXiv:1710.06372*, 2017.

[11] E. Androulaki, A. Barger, V. Bortnikov, C. Cachin, K. Christidis, A. De Caro, D. Enyeart, C. Ferris, G. Laventman, Y. Manevich *et al.*, "Hyperledger fabric: a distributed operating system for permissioned blockchains," in *Proceedings of the thirteenth EuroSys conference*, 2018, pp. 1–15.

[12] S. Shi, D. He, L. Li, N. Kumar, M. K. Khan, and K.-K. R. Choo, "Applications of blockchain in ensuring the security and privacy of electronic health record systems: A survey," *Computers & security*, vol. 97, p. 101966, 2020.

[13] Z. Zheng, S. Xie, H.-N. Dai, W. Chen, X. Chen, J. Weng, and M. Imran, "An overview on smart contracts: Challenges, advances and platforms," *Future Generation Computer Systems*, vol. 105, pp. 475–491, 2020.

[14] H. X. Son *et al.*, "Towards a mechanism for protecting seller's interest of cash on delivery by using smart contract in hyperledger," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 4, pp. 45–50, 2019.

[15] X. S. Ha, T. H. Le, T. T. Phan, H. H. D. Nguyen, H. K. Vo, and N. Duong-Trung, "Scrutinizing trust and transparency in cash on delivery systems," in *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*. Springer, 2020, pp. 214–227.

[16] N. Duong-Trung, H. X. Son, H. T. Le, and T. T. Phan, "Smart care: Integrating blockchain technology into the design of patient-centered healthcare systems," in *Proceedings of the 2020 4th International Conference on Cryptography, Security and Privacy*, ser. ICCSP 2020, 2020, p. 105–109.

[17] ——, "On components of a patient-centered healthcare system using smart contract," in *Proceedings of the 2020 4th International Conference on Cryptography, Security and Privacy*, 2020, p. 31–35.

[18] H. T. Le, K. L. Quoc, T. A. Nguyen, K. T. Dang, H. K. Vo, H. H. Luong, H. Le Van, K. H. Gia, L. V. Cao Phu, D. Nguyen Truong Quoc *et al.*, "Medical-waste chain: A medical waste collection, classification and treatment management by blockchain technology," *Computers*, vol. 11, no. 7, p. 113, 2022.

[19] H. X. Son, T. H. Le, N. T. T. Quynh, H. N. D. Huy, N. Duong-Trung, and H. H. Luong, "Toward a blockchain-based technology in dealing with emergencies in patient-centered healthcare systems," in *International Conference on Mobile, Secure, and Programmable Networking*. Springer, 2020, pp. 44–56.

[20] H. T. Le, L. N. T. Thanh, H. K. Vo, H. H. Luong, K. N. H. Tuan, T. D. Anh, K. H. N. Vuong, H. X. Son *et al.*, "Patient-chain: Patient-centered healthcare system a blockchain-based technology in dealing with emergencies," in *International Conference on Parallel and Distributed Computing: Applications and Technologies*. Springer, 2022, pp. 576–583.

[21] N. T. T. Quynh, H. X. Son, T. H. Le, H. N. D. Huy, K. H. Vo, H. H. Luong, K. N. H. Tuan, T. D. Anh, N. Duong-Trung *et al.*, "Toward a design of blood donation management by blockchain technologies," in *International Conference on Computational Science and Its Applications*. Springer, 2021, pp. 78–90.

[22] H. T. Le, T. T. L. Nguyen, T. A. Nguyen, X. S. Ha, and N. Duong-Trung, "Bloodchain: A blood donation network managed by blockchain technologies," *Network*, vol. 2, no. 1, pp. 21–35, 2022.

[23] H. T. Le *et al.*, "Introducing multi shippers mechanism for decentralized cash on delivery system," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, 2019.

[24] Ethereum, "How our escrow smart contract works," 2022. [Online]. Available: https://www.thenational.ae/business/technology/cash-on-delivery-the-biggest-obstacle-to-e-commerce-in-uae-and-region-1

[25] OpenBazaar, "Truly decentralized, peer-to-peer ecommerce features," 2022. [Online]. Available: https://openbazaar.org/features/

[26] "Two party contracts," 2022. [Online]. Available: https://dappsforbeginners.wordpress.com/tutorials/two-party-contracts/

[27] N. Duong-Trung *et al.*, "Multi-sessions mechanism for decentralized cash on delivery system," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 9, 2020.

[28] N. T. T. Le *et al.*, "Assuring non-fraudulent transactions in cash on delivery by introducing double smart contracts," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 677–684, 2019.

[29] N. M. Hoang and H. X. Son, "A dynamic solution for fine-grained policy conflict resolution," in *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, 2019, pp. 116–120.

[30] H. X. Son and N. M. Hoang, "A novel attribute-based access control system for fine-grained privacy protection," in *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, 2019, pp. 76–80.

[31] Q. N. T. Thi, T. K. Dang, H. L. Van, and H. X. Son, "Using json to specify privacy preserving-enabled attribute-based access control policies," in *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*. Springer, 2017, pp. 561–570.

[32] H. X. Son, T. K. Dang, and F. Massacci, "Rew-smt: a new approach for rewriting xacml request with dynamic big data security policies," in *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*. Springer, 2017, pp. 501–515.

[33] S. H. Xuan *et al.*, "Rew-xac: an approach to rewriting request for elastic abac enforcement with dynamic policies," in *2016 International Conference on Advanced Computing and Applications (ACOMP)*. IEEE, 2016, pp. 25–31.

Enhanced Security: Implementation of Hybrid Image Steganography Technique using Low-Contrast LSB and AES-CBC Cryptography

Edwar Jacinto G, Holman Montiel A, Fredy H. Martínez S
Facultad Tecnológica, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia

***Abstract*—Now-a-days, sensitive and confidential information needs to be exchanged over open, public, and not secure networks such as the Internet. For this purpose, some information security techniques combine cryptographic and steganographic algorithms and image processing techniques to exchange information securely. Therefore, this research presents the implementation of an algorithm that combines the AES-CBC cryptographic technique with the LSB steganographic technique, which is statistically enhanced by image processing by looking for low-contrast areas where the encrypted information will be stored. This hybrid algorithm was developed to send a plaintext file hidden in an image in BMP format, so the changes in the image are invisible to the human eye and undetectable in possible steganographic analysis. The implementation was performed using Python and its libraries PyCryptodome for encryption and CV2 for image processing. As a result, it was found that the hybrid algorithm implemented has three layers of security over a plaintext encrypted and hidden in a digital image, which makes it difficult to break the secrecy of the information exchanged in a stego-image file. Additionally, the execution times of the hybrid algorithm were evaluated for different sizes of plaintext and digital image files.**

Keywords*—Steganography; cryptography; LSB; low contrast areas; AES-CBC algorithm

I. Introduction

When sensitive or confidential information needs to be sent securely between two parties communicating over media with a high probability of attack, e.g., public, open, or unsecured networks such as the Internet, it is necessary to employ information security techniques to perform this exchange. Some techniques can be used for information hiding and others for information encryption [1]. In the case of information concealment, steganographic techniques allow hiding a secret message in a cover message in such a way that its existence is not detectable to others but only to the receiver of the information. In the case of information encryption, cryptographic techniques allow exchanging secret information between sender and receiver through the encryption and decryption of coded messages.

Additionally, when the information traffic on the Internet today is analyzed, it is evident that the conventional type of communication is based mainly on sending images and video, which is how images have come to be selected as a means to communicate secret information securely. Nowadays, there is literature on different image steganographic techniques for information concealment [2] [3], and different cryptographic techniques for information encryption [4]. Such studies classify the existing algorithms, indicate the performance parameters, and show the advantages, possible applications, and attacks or security problems they may present [5] [6].

Thus, some of these studies have concluded that one way to improve information security, increasing the reliability, robustness, and solidity in the exchange of information is to combine steganographic techniques with cryptographic techniques [1] [7]. One way to do this is to take the sensitive or confidential information to be transmitted to perform an encryption process by implementing some cryptographic technique, and then take the encrypted message to perform a mixing process with a cover image using some steganographic technique [8].

Some examples of this are: [9] where the message to be transmitted is encrypted in two stages, the first by Caesar cipher and the second by chaos theory; the encrypted message is embedded in the cover image using the Least Significant Bit (LSB) substitution steganographic algorithm. [10] where the message to be transmitted is encrypted using Advance Encryption Standard (AES) encryption; at the same time, the cover image is preprocessed to resize it and identify the areas where the LSB substitution process was performed using inverse Wavelet Transform and Artificial Neural Networks (ANN). The author in [11] where the message to be transmitted is encrypted using AES encryption, including a hash process; this hash encrypted text is embedded in a cover image through Dynamic Octa Pixel Value Differencing (DOPVD) embedding algorithm that includes LSB + PVD approach. The author in [12] where the image to be transmitted is encrypted using a large secret key through XOR operation; the encrypted image is embedded in a cover image by LSB obtained a stego-image; finally, the stego-image is watermarked in time domain and frequency.

For this reason, this research aims to implement an algorithm that combines cryptographic and statistically enhanced steganography techniques for sending plain text files over a digital image in BMP format. That algorithm develops using Python and the OpenCV libraries as the base implementation language, considering the size restrictions of such information to be hidden as well as the resolution of the cover image [13].

In the case of cryptographic technique, it was decided to use Advanced Encryption Standard (AES) as the encryption

method. Because it is the standard cipher [14] [15] [16], given its security level, information encryption speed (capacity), and current availability in the internal architecture of processors as a dedicated hardware block [17] [18] [19], making it native in any application [20]. The only configurable parameter on AES is the cipher operation mode, which is associated with the order in which the keys and the initialization vector are combined with the information to be encrypted. Therefore, in this case, the Cipher Block Chaining (CBC) operation mode was chosen.

In the case of the statistically enhanced steganographic technique, it was decided to use the Least Significant Bit (LSB) substitution as a base method [21] [22] [23], enhanced in terms of selecting the information hiding areas. Such enhancement is achieved by using image processing techniques to choose a low contrast area [24], where the image entropy is less affected [25], offering a robust solution in terms of a possible stego-analysis. The image processing technique uses applied statistics concepts as mathematical criteria for locating the hidden and encrypted information. For this purpose, it is based on the characteristics of the analyzed images as a random variable, where the histogram's high dispersion can measure an image's high contrast. That is, the higher the contrast of the stego-image, the higher the security level given by this extra layer based on the processing and analysis of digital images (PAID).

Therefore, this paper presents the implementation and validation of a hybrid crypto-steganographic system. Section II describes the structure proposed to implement the hybrid system highlighting the three main elements. Section III explains the development of the software application step by step, showing: how the user key is entered, how the information encryption process, how the area where the encrypted message will be hidden is chosen, and how the execution of the LSB algorithm to reach the output stego-file. Section IV presents the validation of the implemented hybrid system and performs a performance analysis of the complete application's processing time. Finally, in Section V the conclusions according to these results are shown.

II. Methodology

The proposed structure is a hybrid technique that combines the AES-CBC cryptographic technique with enhanced LSB steganography to hide the information in the lower contrast area of the image, as shown in Fig. 1. The upper part of the graph shows the plaintext encryption process corresponding to the sensitive or confidential information to be transmitted. The lower part shows the processing of the cover image to determine the low contrast area. The right side of the figure shows the embedded process of the cipher text using a classical LSB technique to conceal it in the low contrast area, obtaining a stego-file containing the encrypted information immersed in the image (called Stego-Image).

This proposed structure complies with the philosophy of Feistel networks, which is none other than having the same architecture of the solution for the encryption and decryption of the information. It is a reversible structure where it is only necessary to reverse the order of the blocks to carry out the decryption process.

For the implementation, an application was made in Python 3.X, using the OpenCV libraries for image processing, in addition to using Numpy to work with vectors and matrices and MatplotLib to visualize the partial and final results. In this case, the information to be encrypted and hidden is a plain text encoded in UTF-8, to which a data type change process must be performed to be encrypted with AES in a CBC operation mode, always working with a pure binary string or in 64 bits format. Finally, the Cryptodome library was used, which has every one of the cryptographic functions necessary to encrypt the information.

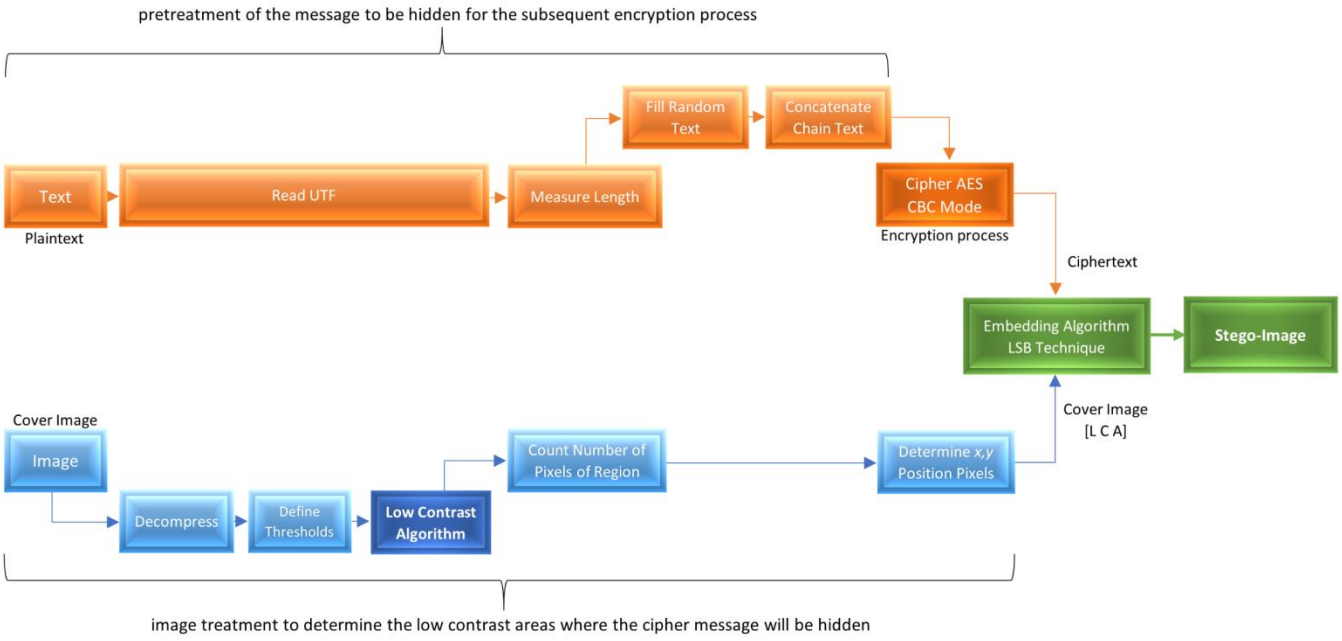


Fig. 1. General Block Diagram Implementation.

A. User Key Expansion

The aim is to encrypt the plain text using a standard block cipher such as AES-CBC; this algorithm requires a key and an initialization vector. For that, a key expansion process must be performed, and then the encryption of this key with an asymmetric algorithm [26].

Since block ciphers must have a key size equal to the block's size to be encrypted, a minimum procedure is required to ensure that the size of the session key is entered by the person or entity that will perform the steganography process complies with this characteristic. For this purpose, a key expansion procedure is performed, using the functions of the Cryptodome library developed for Python in its version 3.0 or higher, employing the PKCS #5 key expansion process, shown in the pseudo-code of Fig. 2.

This block cipher algorithm generates a series of subkeys to increase the entropy of the encrypted information; this process of generation and mixing of the subkeys with the information to be encrypted corresponds to the CBC mode of operation. This mode of operation generates high entropy since, in each of the cipher rounds, the K-th subkey is mixed with the information, obtaining at the end of the encryption process a new key that is mixed with the initialization vector. This process is automated and standardized, achieving greater security in the encryption of the information.

B. Encryption Process of the Plaintext Message

Once the operation mode of the encryptor, its working mode, the key of the indicated size, and the implicit generation of the initialization vector are clear, the information encryption process must be clearly understood to achieve its compatibility with the image file. In this case, the size of the plaintext file must be clear to be encrypted, what type of data is necessary to achieve the encryption process, and generate empty vectors where the encrypted information will be returned. For this process, it must have some of the tools offered by the NumPy library.

Next, the process of reading the text size to be encrypted is performed and compared with the number of pixels chosen to blend the ciphertext with the image. It must be guaranteed that the number of characters of the input text will never be the same as the number of pixels; therefore, a random filler text must be generated so that no empty spaces are created in the output stego-file. This process is done to avoid the simple detection of hidden information. Fig. 3 shows in a simple pseudo-code this programming scheme.

Once the character string is ready to perform the information encryption process, taking into account the dimensions and type of data required, an algorithm is applied that selects the low contrast areas where it is more difficult to detect the hidden information.

C. Statistical Method: Choice of the Low Contrast Area

The aim is the cover image file processing using applied statistics concepts to identify the areas of lower contrast in the image, areas where the encrypted information will be placed. In other words, a process of selecting in which part of the image to place the hidden information is made. This task requires using image processing functions available in the OpenCV (CV2) libraries.

According to [24], an algorithm is used to detect some pixels with low contrast, complying with the following criteria to detect the areas where the human eye does not detect any change:

The criterion for detecting dark areas is in the image. This criterion is described in (1). It requires determining a working window, the number of pixels of the convolution matrix. For establishing a series of local medians $m_{s_{xy}}$, which will be compared with the global median of the entire image m_G using a weighting constant k_0, where the value of this constant depends on the gray level that will be given as "dark."

$$m_{s_{xy}} \leq k_0 \cdot m_G \quad (1)$$

```
1.  Input = KeySession //the user enters the key
2.  Salt = str(rnd(N_std)) //generate the random number
3.  for I in 0 to N_times
4.      AES_KEY = HASH (Input + Salt) // calculate the digest N times to the
5.  End
```

Fig. 2. Pseudo-code for user Key Expansion.

```
1.  PlainText = open (AnyFile.txt.encode(UTF-8)) // Open File
2.  SizeText = len(PlainText*8) //8 bits per character  // calculate length per char to pixel
3.  ImgPixels = CountPixels (rows*columns)  //measure length of the image
4.  Bits_Array = zeros(ImgPixels) // create an empty array
5.  Len_Random_str = round (ImgPixels - SizeText / 8) //calculate length of the fill text
6.  For I in range (0, len_random_str):
7.          Random_str[i] = random('a','z') //generate random char
8.  End
9.  Plain_text_fill = plain_text +random_chain //create the final char array
10. Plain_Text_Bytes = str.encode.bytes(plain_text_fill) // To Cypher plaint text in bytes only
```

Fig. 3. Pseudo-code for Encryption Process of the Plaintext Message.

The possible selection criterion is for finding low contrast areas. This criterion is described in (2). It compares the local standard deviation in a certain pixel window $\sigma_{s_{xy}}$ with the global standard deviation σ_G of the whole image by taking as a weighting factor or comparison criterion a factor k_2. This factor is determined by the experience of how scattered are the grayscale values in the low contrast regions to be detected.

$$\sigma_{s_{xy}} \leq k_2 \cdot \sigma_G \quad (2)$$

On the other hand, a possible error is generated in the selection criterion that [24] describes as enhancing a constant area, where the standard deviation would be zero evidently. Such a problem must be applied depending on the characteristics of the chosen image and is described by (3).

$$K_1 \cdot \sigma_G \leq \sigma_{s_{xy}} \quad (3)$$

This equation describes the way to compare a minimum local standard deviation in a certain pixel window $\sigma_{s_{xy}}$ with the global standard deviation σ_G with a factor K_1, avoiding enhancing or selecting constant zones. In other words, it becomes undesirable to select a pixel from a zone with the same gray level as a candidate for the LSB algorithm.

The algorithm applying the above mathematical criteria in practice was implemented through a pseudo-code, shown in Fig. 4. As a result, a binary matrix is obtained, which clearly identifies the low contrast zones where the LSB information mixing algorithm will be used to obtain the stego-image with the concealed information in these specific zones.

D. LSB Method (Least Significant Bit)

It is a method that seeks to place a binary string with the information to be hidden in the stego-file. In this case, having encrypted information, the binary string will be in a pure binary format or base64; these types of data result from the encryption process. Then, mixing or embedding this information in the cover image is performed through a simple binary mask. Fig. 5 shows the pseudo-code that mixes or embeds the encrypted information in the least significant bit of the cover image in the areas chosen by the statistical algorithm (low contrast areas).

It is necessary to ensure the correct functioning of the LSB algorithm that an image with large low-contrast areas compared to the total image size should be chosen. On the other hand, it is recommended to work with images of a size larger than the possible size of the plaintext to be encrypted and hidden in the digital image. In other words, the size of the stego-image of the input image should be much larger than the stego-message, which should be cipher using an encryption algorithm.

For this case, it was chosen only to hide plaintext files since it requires less processing than applying such processing to multimedia files. However, if the application requires it, the same technique can be applied to other types of files or combinations.

```
1.  Img = cv2.read(ImFile.Imext) //read the image
2.  ImgGray = cv2.cvtColor(Img, cv2.COLOR_BRG2GRAY) //Convert to only one matrix
3.  GlobalMedian = np.mean(ImgGray)  // calculate global median
4.  GlobalDevStd = np.std(ImgGray) // calculate global standard deviation
5.  Pix = NWin
6.  Dim = (pix,pix)
7.  MConv = np.ones(Dim) * 1/(pix**2)  // create convolution matrix
8.  LocalMedian = conv2(ImgGray,Mconv) // calculate local median
9.  LocalDevStd = conv2((ImgGray - LocalMedian) .^2 , MConv))  // calculate local standard deviation
10. ImgOut1 = LocalMedia <= K0*GlobalMedian  // compare to create binary matrix of dark regon
11. ImgOut2 = LocalDesvStd <= k2* GlobalDesvStd // compare to create binary matrix of low contrast
12. ImgOut = BitAND(ImgOut1, ImgOut2) // Binary matrix of the choose pixels
```

Fig. 4. Pseudo-code for Statistical Method - Choice of Low-contrast Area.

```
1.  Def Func_LSB(msg): // define LSB function
2.      for CharIn in(msg):
3.          o = CharIn  //// read the encrypted message
4.          for Column in (Img):  // traverse the columns
5.              for Row in (Img): // traverse the rows
6.                  if ImgOut(Row,Column) == (TRUE) // check if the pixel is choose to lsb
7.                      for in range (8)
8.                          O & lsb(Img) // put the information in the 8 pixels
```

Fig. 5. Pseudo-code for LSB Method.

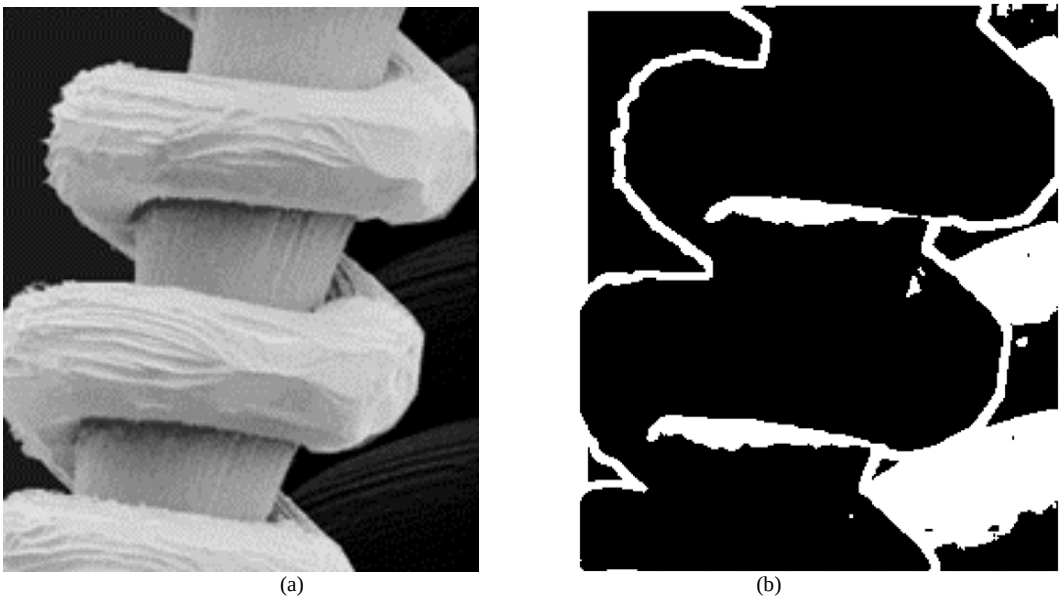
III. Results

The first step to verify the algorithm's effectiveness is to verify that the statistical algorithm effectively identifies the areas with low contrast. Fig. 6 shows in part (a) the original image with an area of low contrast and in part (b) a black and white image, where the white parts are the areas of the pixels chosen to perform the information hiding process using the LSB algorithm. For this part, it was only necessary to follow the steps of equations (1) and (2).

As a result, the stego-image does not look the same in its least significant bits as the original cover image, although this is not as visible to the human eye. The image loses its natural entropy, i.e., the bright part of the image is removed. This characteristic serves as an indication to discover whether a stego-image has hidden information in a steganalysis process.

The information encryption and hiding process were tested using the AES-CBC algorithm for encryption and the enhanced LSB algorithm that mixes o embeds information in the low contrast areas; this test was made using different sizes of plaintext files. Fig. 7 shows a bar graph that presents the execution time depending on the size of the information to be encrypted and hidden. The tests were performed on a PC with an 8-core Core i7 with 16 Gigabytes of RAM and a Geforce GTX 610 video card.

It can be seen how the application can store files of different sizes in the stego-image, up to a limit of one Megabyte, the size of the book Don Quixote in plaintext format, for which the processing time was approximately one minute. On the other hand, the time for information smaller than 200 kilobytes is less than 10 seconds, so it could be said that the process is agile for small texts.



(a) (b)

Fig. 6. (a) Original Image. (b) Image with Low-contrast Areas in White [24].

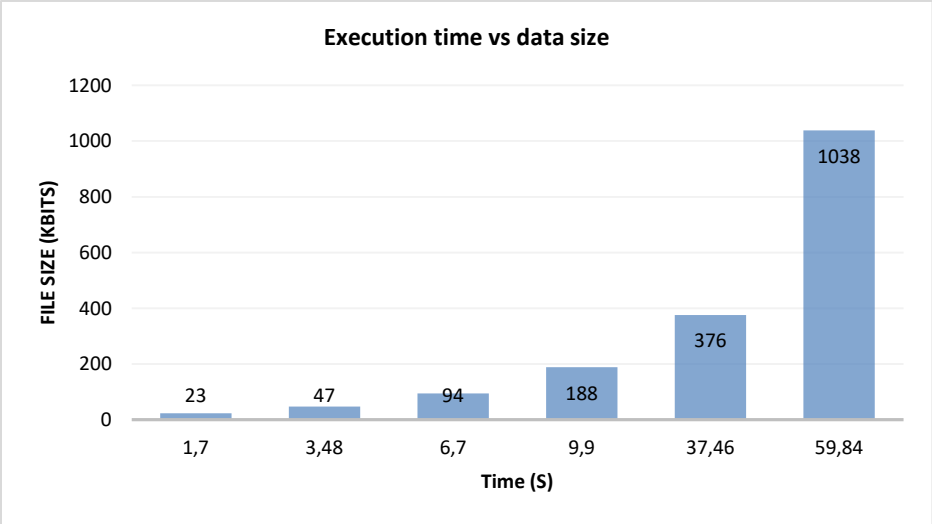


Fig. 7. Performance Graph of the Algorithm Measured on Text Files of Different Sizes.

IV. Conclusion

It was verified that the step-by-step implementation of a hybrid algorithm that combines cryptographic and steganography techniques for sending plaintext files over a digital image in BMP format gives a double layer of security. So, if the stego-image is revealed to have hidden information, it is impossible for the person or entity intercepting the message to know plaintext content because it is encrypted.

On the other hand, it was determined that using a statistically enhanced steganography technique to choose the lower contrast areas to hide the encrypted information only in these zones gives the hybrid algorithm an extra layer of security, making the entire algorithm more robust. The choosing lower contrast areas algorithm makes that the entropy of the image is only affected in the areas chosen by it. This feature makes it difficult to detect concealed information. It adds an extra layer of security since, besides having the session key to decrypt the ciphertext, the value of two constants, k_0 and k_2, must be present so that when performing the decryption process, the information only is taken from these areas. Therefore, it is verified that only modifying the least significant bit does not affect the statistical selection criteria with which the pixels in which the encrypted information was hidden were chosen.

Regarding the cryptographic technique implemented to encrypt the information before hiding it, the standardized AES-CBC algorithm was used, which was automated using the Cryptodome library, achieving greater security in the encryption of the information. However, it became evident that it would be possible to experiment with different combinations of standardized modes for AES in the Cryptodome library for future work. Seeking to maximize the entropy in the information and therefore generate fewer possible patterns in the LSB algorithm, as well as specifying how the key and the comparison constants would be exchanged.

Finally, the algorithm's performance was analyzed regarding the time used for the encryption and embedded process, resulting in fast usability for small plaintext files below 200 kilobytes. It is a good performance considering that a 100 kilobytes text is the entire chapter of any chapter literature text.

Acknowledgment

The Universidad Distrital Francisco José de Caldas supports this work through the research group SIE -Embedded Informatics Security- which belongs to the Technological Faculty. SIE has dedicated to working in cryptography and applied steganography. Currently, the bases are being generated to implement this type of algorithm in stand-alone applications, which is the final purpose of the workgroup.

References

[1] M. S. Taha, M. S. Mohd Rahim, S. A. Lafta, M. M. Hashim, and H. M. Alzuabidi, “Combination of Steganography and Cryptography: A short Survey,” in IOP Conference Series: Materials Science and Engineering, Jun. 2019, vol. 518, no. 5. doi: 10.1088/1757-899X/518/5/052003.

[2] A. O. Vyas and S. v Dudul, “An Overview of Image Steganographic Techniques,” International Journal of Advanced Research in Computer Science, vol. 6, no. 5, pp. 67–72, 2015, Accessed: Jun. 30, 2022. [Online]. Available: http://www.ijarcs.info/index.php/Ijarcs/article/viewFile/2483/2471

[3] S. G. Shelke and S. K. Jagtap, “Analysis of spatial domain image steganography techniques,” in Proceedings - 1st International Conference on Computing, Communication, Control and Automation, ICCUBEA 2015, Jul. 2015, pp. 665–667. doi: 10.1109/ICCUBEA.2015.136.

[4] G. C. Kessler, “An Overview of Cryptography,” Jun. 2010. [Online]. Available: www.garykessler.net/library/crypto.html.

[5] D. Laishram and T. Tuithung, “A Survey on Digital Image Steganography: Current Trends and Challenges,” May 2018. Accessed: Jun. 30, 2022. [Online]. Available: https://ssrn.com/abstract=3171494.

[6] B. Jana, M. Chakraborty, T. Mandal, and M. Kule, “An Overview on Security Issues in Modern Cryptographic Techniques,” May 2018. Accessed: Jun. 30, 2022. [Online]. Available: https://ssrn.com/abstract=3173527.

[7] A. Jan, S. A. Parah, M. Hussan, and B. A. Malik, “Double layer security using crypto-stego techniques: a comprehensive review,” Health and Technology, vol. 12, no. 1, pp. 9–31, Jan. 2022, doi: 10.1007/s12553-021-00602-1.

[8] S. Almuhammadi and A. Al-Shaaby, “A Survey on Recent Approaches Combining Cryptography and Steganography,” in Computer Science & Information Technology (CS & IT), Feb. 2017, pp. 63–74. doi: 10.5121/csit.2017.70306.

[9] G. S. Charan, Nithin Kumar S S V, Karthikeyan B, Vaithiyanathan V, and Divya Lakshmi K, “A novel LSB based image steganography with multi-level encryption,” in 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Mar. 2015, pp. 1–5. doi: 10.1109/ICIIECS.2015.7192867.

[10] K. S. Seethalakshmi, Usha B A, and Sangeetha K N, “Security enhancement in image steganography using neural networks and visual cryptography,” in 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Oct. 2016, pp. 396–403. doi: 10.1109/CSITSS.2016.7779393.

[11] S. Mangela, N. Daddikar, T. Bargode, and P. N. Tatwadarshi, “Advance steganography using dynamic octa pixel value differencing,” in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Mar. 2017, pp. 1–7. doi: 10.1109/ICIIECS.2017.8275989.

[12] M. Abdur, R. Ahmed, M. Adnan, and A. Ahmed, “Digital Image Security: Fusion of Encryption, Steganography and Watermarking,” International Journal of Advanced Computer Science and Applications, vol. 8, no. 5, 2017, doi: 10.14569/IJACSA.2017.080528.

[13] T. Morkel, “Self-sanitization of digital images using steganography,” in 2015 Information Security for South Africa (ISSA), Aug. 2015, pp. 1–6. doi: 10.1109/ISSA.2015.7335073.

[14] P. Sethi and V. Kapoor, “A Proposed Novel Architecture for Information Hiding in Image Steganography by Using Genetic Algorithm and Cryptography,” Procedia Computer Science, vol. 87, pp. 61–66, 2016, doi: 10.1016/j.procs.2016.05.127.

[15] M. E., A. A., and F. A., “Data Security Using Cryptography and Steganography Techniques,” International Journal of Advanced Computer Science and Applications, vol. 7, no. 6, 2016, doi: 10.14569/IJACSA.2016.070651.

[16] R. Indrayani, H. A. Nugroho, R. Hidayat, and I. Pratama, “Increasing the security of mp3 steganography using AES Encryption and MD5 hash function,” in 2016 2nd International Conference on Science and Technology-Computer (ICST), Oct. 2016, pp. 129–132. doi: 10.1109/ICSTC.2016.7877361.

[17] I. Algredo-Badillo, F. R. Castillo-Soria, K. A. Ramírez-Gutiérrez, L. Morales-Rosales, A. Medina-Santiago, And C. Feregrino-Uribe, “Lightweight Security Hardware Architecture Using DWT and AES Algorithms,” IEICE Transactions on Information and Systems, vol. E101.D, no. 11, pp. 2754–2761, Nov. 2018, doi: 10.1587/transinf.2018EDP7174.

[18] V. Sharon, B. Karthikeyan, S. Chakravarthy, and V. Vaithiyanathan, “Stego Pi: An automated security module for text and image steganography using Raspberry Pi,” in 2016 International Conference on Advanced Communication Control and Computing Technologies

(ICACCCT), May 2016, pp. 579–583. doi: 10.1109/ICACCCT.2016.7831706.

[19] A. Odeh, K. Elleithy, and M. Faezipour, "Fast real-time hardware engine for ZWC text steganography," in 2014 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE), May 2014, pp. 1–5. doi: 10.1109/VITAE.2014.6934454.

[20] D. Das and M. Dutta, "Security Enhancement on Application Oriented Steganographic Schemes with Crypto-Encryption: A Technical Review," TIU Transactions on Intelligent Computing (TTIC), vol. IV, Dec. 2020, [Online]. Available: https://www.researchgate.net/publication/358280401.

[21] P. Martí Méndez Naranjo and D. Fernando Avila Pesantez, "Cryptography application experience to improve security in a steganographic method in images," Revista Espacios, vol. 40, no. 38, Nov. 2019, Accessed: Jul. 01, 2022. [Online]. Available: https://www.researchgate.net/publication/350153611.

[22] S. Raniprima, B. Hidayat, and N. Andini, "Digital image steganography with encryption based on rubik's cube principle," in 2016 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), Sep. 2016, pp. 198–201. doi: 10.1109/ICCEREC.2016.7814972.

[23] S. L. Chikouche and N. Chikouche, "An improved approach for lsb-based image steganography using AES algorithm," in 2017 5th International Conference on Electrical Engineering - Boumerdes (ICEE-B), Oct. 2017, pp. 1–6. doi: 10.1109/ICEE-B.2017.8192077.

[24] R. C. Gonzalez and R. E. Woods, Digital Image Processing, Third. Pearson, 2008.

[25] R. Roy and S. Changder, "Image steganography with block entropy based segmentation and variable rate embedding," in 2014 2nd International Conference on Business and Information Management (ICBIM), Jan. 2014, pp. 75–80. doi: 10.1109/ICBIM.2014.6970937.

[26] E. J. G. et al. , Edwar Jacinto Gómez et al., "Implementation of a Crypto-Steganographic System Based on the Aes-Cbc Algorithm," International Journal of Mechanical and Production Engineering Research and Development, vol. 10, no. 3, pp. 15059–15068, Jun. 2020, doi: 10.24247/ijmperdjun20201435.

Secure and Efficient Implicit Certificates: Improving the Performance for Host Identity Protocol in IoT

Zhaokang Lu[1]
School of Computer Science and Technology
Harbin University of Science and Technology
Harbin, 150080, China

Jianzhu Lu[2*]
Department of Computer Science
Jinan University
Guangzhou, 510630, China

***Abstract*—Implicit certificates own the shorter public key validation data. This property makes them appealing in resource-constrained IoT systems where public-key authentication is performed very often, which is common in Host Identity Protocol (HIP). However, it is still a critical challenge in IoT how to guarantee the security and efficiency of implicit certificates. This article presents a forgery attack for the Privacy-aware HIP (P-HIP), and then propose a Secure and Efficient Implicit Certificate (SEIC) scheme that can improve the security of the P-HIP and the efficiency of elliptic-curve point multiplications for IoT devices. For a fix-point multiplication, the proposed approach is about 1.5 times faster than the method in SIMPL scheme. Furthermore, we improve the performance of SEIC with the butterfly key expansion process, and then construct an improved P-HIP. Experimental results show that compare to the existing schemes, the improved scheme makes a user/device have both the smallest computation cost and the smallest communication cost.**

***Keywords*—Authentication; privacy; implicit certificates; internet of things (IoT); host identity; security**

I. Introduction

Public-key authentication is a critical issue for any IoT system. Many current IoT devices rely on the Public Key Infrastructure (PKI) to achieve public-key authentication [1]. One of the differences between IoT devices and conventional devices is how their public-keys are authenticated. Traditional devices rely basically on X.509 certificates [2] whereas IoT devices use implicit certificates [3]. It is a known fact that besides the device's public key and the signature generated by a trusted Certificate Authority (CA), an X.509 certificate also contains information about the certificate subject, the supported encryption and/or digital signing algorithms and information to determine the revocation and validity status of the certificate. For instance, for the authentication and key exchange of HIP shown in Fig. 1, the certificates $cert_r$ and $cert_r$ are carried in the messages R_1 and I_2, respectively. Digital signatures of the parties are applied for this situation. In order to verify the correctness of a message signed by the private key of the sender, a receiver first needs to validate the corresponding public-key via its certificate. When IoT devices use implicit certificates instead of explicit certificates, they will reduce the amount of information required for authenticating public keys. Hence, public-key authentication in IoT encourages the use of implicit certificates.

This work focuses on a specific yet important problem: how to attain fast public-key authentication in IoT through the use of Elliptic Curve (EC) implicit certificates. Allowing one IoT device to achieve public key validation remains a challenging problem [1]. This is because certificate verification involves expensive public key operations and communication cost transmitting the authentication data. From Table I, we notice that the certificate size is much larger than the public key size, and their ratio is greater than 5. Here, the certificates are self-signed and created using the OpenSSL library [2]. Some authors [3, 4] explored how to reduce the communication overhead for public-key authentication through the use of EC implicit certificates. However, the scheme in [3] suffers from a forge attack (see Section II-B). This paper will propose a secure scheme to further improve the efficiency of public key authentication by using EC implicit certificates. That is, by designing a table-and-optimality-based technique for EC fixed-point multiplication, we improve the computational efficiency of public key authentication. In addition, by integrating the proposed scheme with butterfly key expansion process, we also reduce the communication cost of public key authentication.

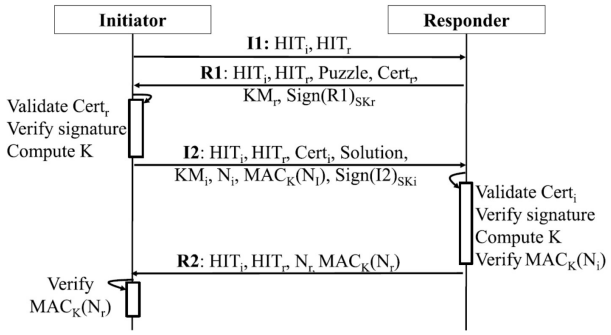


Fig. 1. HIP Handshakes for Authentication and Key Exchange.

TABLE I. X.509 Certificate Size Versus Public-Key Length

RSA				ECC			
KL	KS_1	CS	rat	KL	KS_2	CS	rat
1024	128	1237	9.66	112	28	948	33.85
2048	256	1590	6.21	160	40	981	24,52
3072	384	1935	5.03	256	64	1050	16.40
KL: Key Length(bits); KS_1: Modulus N size(Bytes); CS: Certificate Size(Bytes)							
KS_2: Uncompressed point Q(X,Y) size(Bytes); rat:CS/KS_1 or CS/KS_1							

Fast public-key revocation validation seems to be a dilemma on IoT devices. In PKI, public-key revocation is achieved through the certificate revocation. The unexpired

* Corresponding author

certificates are revoked by using Certificate Revocation Lists (CRLs) or the Online Certificate Status Protocol (OCSP). The CRL introduces substantial communication overhead since the CRL size is proportional to the number of revoked certificates. The OCSP increases certificate revocation verification latency and the risk of leaking user privacy (such as accessing history of the device). Hence none of the above methods is desired for IoT. In order to reduce both the communication overhead and the latency, based on Nyberg's one-way accumulator, this paper designs a credential revocation mechanism which is significantly efficient in the verifier's side.

With the applications of IoT technology, security and privacy concerns have also emerged. Especially, tracking of devices' activities is a threat to their privacy. HIP is a suitable solution for IoT devices considering the security and privacy requirements of IoT systems [3]. In the HIP, an IoT device is issued a public key as host identifier and a 128-bits hash of the public key as Host Identity Tag (HIT). A mobile IoT device uses the same public key and HIT to authenticate to its peers when it moves from one network to another network. By learning the public key and HIT, an attacker can track the mobile IoT device. This paper improves the security of P-HIP in [3] such that devices can avoid tracking by changing the short-lived implicit certificates.

In general, the different implementations of public-key authentication in IoT should at least fulfill the following requirements: (1)Security: The CA can not be compromised or coerced to assign a public key to a malicious attacker. (2)Accuracy: A device should accurately determine a certificate revocation status. (3)User privacy: The protocol should not leak the identities of the accessing devices. (4)Efficiency: The protocol should cost small memory, computation, and network resource on IoT devices. (5)Compatibility: The protocol is required to be compatible with current certificate standards and existing certificates.

Our Contribution: In this paper, we first present a forgery attack for P-HIP scheme [3], and then propose a secure and efficient implicit certificate (SEIC) scheme to overcome its weaknesses. Specifically, the SEIC scheme runs the signature algorithm to output a signature by hashing the public key validation data, the timestamp and the CA's public key. A table-and-optimality-based technique is designed for EC fixed-point multiplication such that it's achievement is about 1.5 times faster than the method [4]. In addition, SEIC contains a credential revocation mechanism which is significantly efficient in the verifier's side. That is, the verifier achieves the revocation verification of a public key by performing one Nyberg one-way accumulator operation while keeping only one $\mathbb{Z}_p$ symbol. Furthermore, we improve SEIC via the butterfly key expansion process, and then construct an improved P-HIP. Experimental results show that the improved P-HIP can achieve performance gains during credential issuance and mutual authentication, while preserving the user privacy.

The rest of this paper is organized as follows: Section II analyzes the security of the implicit certificate scheme in P-HIP [3] after an overview of the scheme, and then introduces related work. Section III explains the basics notations as well as the primitives of proposed scheme. Section IV introduces SEIC scheme and discusses its security and performance. Section V shows that SEIC can be improved via the butterfly key expansion process, and then an improved P-HIP is constructed. Section VI formally analyzes the privacy of the improved P-HIP and the corresponding performance gains is shown in Section VII. Finally, Section VIII concludes the discussion.

user	CA	user
	20. $D_{d_{ca}}(E_{Q_{ca}}(R_u))$	31. $\alpha_u = r_u * Q_{ca}$
	21. $k_u \in [n-1]$	32. $D_{\alpha_u}(E_{\alpha_u}[s_u, \delta_u])$
11. $r_u \in [n-1]$	22. $\delta_u = k_u * G$	33.Verify $Mac_{\alpha_u}[s_u \| \delta_u]$
12. $R_u = r_u * G$	23. $s_u = k_u + d_{ca}$	34. $d_u = r_u + s_u \bmod n$
$\xrightarrow{E_{Q_{ca}}(R_u)}$	24. $\alpha_u = d_{ca} * R_u$	35. $Q_u = d_u * G$
		36. $V_u = \delta_u + R_u$
	$\xrightarrow{E_{\alpha_u}[s_u, \delta_u], Mac_{\alpha_u}[s_u \| \delta_u]}$	37. $Q_u =^? V_u + Q_{ca}$
		38.Store s_u, V_u, d_u, Q_u
		39.destroy r_u, R_u, δ_u

Fig. 2. The Implicit Certificate Scheme in P-HIP [3].

II. Analysis of the Implicit Certificate Scheme in P-HIP and Related Work

A. The Implicit Certificate Scheme in P-HIP

Consider an additive cyclic group $\mathcal{G}$ generated by a point G on the elliptic curve $y^2 = x^3 + ax + b$ over a finite field $\mathbb{F}_q$, where q is a large prime and $4a^3 + 27b^2 \neq 0 (\mathrm{mod} q)$, and n is the order of generator G. We assume that Q_{ca} and $Q_{\overline{ca}}$ in $\mathcal{G}$ are the public keys of CA and $\overline{\mathrm{CA}}$, respectively.

As shown in Fig. 2, an implicit certificate scheme was proposed recently to design a P-HIP [3]. We review the scheme as follows. The basic goal of the scheme was to bind a public key Q_u to its owner u via the public-key authentication data V_u. To compute the private key construction data s_u, the scheme is different from the conventional ECQV implicit certificate schemes. That is, the CA does not issue a certificate ($cert_u$) to the user. The scheme computed s_u as $s_u = k_u + d_{ca}(modn)$, which did not multiply h_u=$Hash(cert_u)$ with k_u or d_{ca} to compute s_u (see the steps 23 of Fig. 2). Then, s_u and δ_u were encrypted and then sent to the user. Upon receiving a new ECQV-based credential s_u and δ_u, the user computed a unique public key Q_u=$d_u * G$ and HIT HIT_u=$Hash(Q_u)$ for a network that it would join without communicating with the CA.

A user device provided its ECQV public key Q_u and the public-key authentication parameter V_u to a verifier device. The verifier computed a public key Q'_u as Q'_u=$V_u + Q_{ca}$. If Q'_u=Q_u, then the verifier ensured that the public key was genuine and issued by the CA.

B. Security Weaknesses of the Implicit Certificate Scheme in P-HIP

We have observed weaknesses of the implicit certificate scheme in P-HIP [3]. First, the scheme suffered from a forgery attack. This is because a malicious user holding the implicit certificate issued by a CA is able to forge an implicit certificate issued by another CA, precluding the use of any digital signature scheme. In the case, as shown in Fig. 2, the CA whose public key is Q_{ca} had issued an implicit certificate (V_u, d_u, Q_u) to u. Assume that an adversary either is u

himself or colluding with him. Let $\overline{\text{CA}}$ be compromised by the adversary who knows its public key $Q_{\overline{\text{CA}}} = (\overline{x}, \overline{y})$. The adversary disguising as $\overline{\text{CA}}$ can mount a forge attack such that a forged $(\overline{V}_u, d_u, Q_u)$ passes in the public key verification. Here, $\overline{V}_u = V_u + Q_{ca} + (-Q_{\overline{ca}})$, and $-Q_{\overline{ca}} = (\overline{x}, -\overline{y} \bmod n)$. This is because in the scheme [3], $\overline{\text{CA}}$ had issued the public key Q_u to u if and only if its implicit certificate satisfied the equation $Q_u = \overline{V}_u + Q_{\overline{ca}}$. It is easy to see that the equation holds under the condition that $Q_u = V_u + Q_{\underline{ca}}$. This means that the forged implicit certificate under the $\overline{\text{CA}}$ is valid to u. Therefore, the implicit certificate scheme in P-HIP [3] is insecure.

In addition, we deplore implicit certificate schemes in [3, 4] lack of a revocation mechanism. Thus, there is a risk that a malicious attacker might try to use the relevant credential that it is no longer valid while the credential itself has not expired. Instead of CRLs and OCSP, we propose a certificate revocation mechanism that uses only one Z_p-symbol as the authentication information to achieve the revocation validation of implicit certificates.

C. Related Work

Public-key authentication is emerged as a popular tool in IoT applications. In PKI, the CA issues and manages public keys of users by using digital certificates. In the traditional explicit certification model,a user's digital certificate $cert_u = (meta, Q_u, sig_u)$ is issued by a trusted CA. The signature sig_u on $cert_u$ implies that the owner of $cert_u$ knows the private key d_u of public key Q_u. In the implicit certification model, the key pair (d_u, Q_u) is computed by the user u in collaboration with the CA. Implicit certificates were introduced in the work of Günther [5] and Girault [6]. Brown et al. [7] defined a general notion of security for implicit certificates, and proved that optimal mail certificates were secure under this definition. However, it has various drawbacks in terms of security and efficiency. In 2013, Campagna [8] presented an implicit certification solution in the Elliptic Curve Qu-Vanstone (ECQV) protocol. Unfortunately, this approach suffered from certificate misbinding attacks. Recently, Barreto et al. [4] proposed an improvement for its security weaknesses and computational efficiency.

The authenticated key establishment between two IoT devices was achieved via HIP [9]. Fig. 1 shows that the host identifiers (public keys and HITs) were validated by the HIP peers exchanging X.509 certificates. However, the size of the certificate is much larger than both that of its public key (see Table I) and the maximum transmission unit of the IEEE 802.15.4 link [10] in IoT networks. Recently, Hossain and Hasan [3] proposed P-HIP in which the ECQV implicit certification scheme was able to reduce the public-key authentication data for mutual authentication while protecting the user privacy. In this work,we shows that the ECQV implicit certificate in P-HIP suffered from a forgery attack, that is, a malicious user holding the implicit certificate issued by a CA was able to forge an implicit certificate issued by another CA. Then, a new scheme SEIC is proposed to resist the forge attacks.

III. Preliminaries

In this section, we introduce some notations and Nyberg's one-way accumulator needed later.

A. Notations

We shall use the following notations throughout the paper. A set with integers $1, 2, \cdots, n-1$, is written either $\mathbb{Z}_n^*$ or simply $[n-1]$. We denote by $|x|$ the length of the binary string corresponding to x, and $\lceil x \rceil$ the least integer that is greater than or equal to the given number x. Let $\mathbb{F}_q$ be a finite fields, $\mathbb{Z}_n$ be a addition group, and $\mathbb{Z}_n^* = \mathbb{Z}_q \setminus \{0\}$, where q and n are two prims, $q \geq n+1$, and n is the size of a signature (see step 24 in Fig. 3 and 5). We let $H : \{0,1\}^r \times \{0,1\}^* \to \{0,1\}^r$ denote a Nyberg one-way accumulator, $Hash : \{0,1\}^* \to \mathbb{Z}_n$ and $h : \{0,1\}^* \to \{0,1\}^{rd}$ be two one-way hash functions, where h is used to construct the required H. Let p is a prime number satisfying r=$|p|$, where $A_{\sqcup} \in \mathbb{Z}_p$ (see Section IV-B2)).

B. Nyberg's One-Way Accumulator

Here, the concept of the Nyberg One-Way Accumulator (NOWA) in [11] is reviewed. Let $H(\cdot,\cdot)$ denote NOWA from $\{0,1\}^r \times \{0,1\}^*$ to $\{0,1\}^r$, and $\odot$ be the bitwise operation AND. The NOWA was constructed by a one-way hash function $h : \{0,1\}^* \to \{0,1\}^{rd}$. Here, $N = 2^d$ is an upper-bound to the number of items to be accumulated, and $r = |q|$ is an integer. All that was required to specify an NOWA was hashing process and AND operation.

Let $h_1, h_2, \cdots, h_n$, $n \leq N$ be the items to be accumulated, and $h(h_i) = y_i$, $i = 1, \cdots, m$ be their hash values. Each hash value is a string of length rd bits. The heart of NOWA was the hashing process. The hashing process applied a hash function h to the input to produce a r-bit output. The hashing process was composed of the following operations: (1)Hashing operation: Hash accumulated item h_i of the input and output a rd bits binary string v_i=$h(h_i)$. (2)Transfer α: NOWA did a transfer operation on the binary string v_i which was divided into r blocks, $(v_{i,1}, \cdots, v_{i,r})$, of length d. The transfer of a block from a d-bit input to a bit output was performed as follows: If $v_{i,j}$ was a string of zero bits, it was replaced by 0; otherwise, $v_{i,j}$ was replaced by 1. That is, $\alpha(v_i)$ =$(b_{i,1}, \cdots, b_{i,r})$, where $b_{i,j} \in \{0,1\}$, j=$1, \cdots, r$. In this way, we can transfer an accumulated item h_i to a bit string, b_i=$\alpha(h(h_i)) \in \{0,1\}^r$, which can be considered as a value of r independent binary random variable if h is an ideal hash function.

In practice, the NOWA is effectively implemented by using the generic symmetry-based hash function and simple bit-wise operations. The NOWA on an accumulated item $h_i \in S$ with an accumulated key $k \in \{0,1\}^r$ was able to be implemented using the AND operation described as $H(k, h_i) = k \odot \alpha(v_i) = k \odot \alpha(h(h_i))$. And it also could be represented as $A = H(k, h_i) = k \odot \alpha(v_i) = k \odot \alpha(h(h_i))$ $(i \in [n])$ if S was a set of accumulated items $S = \{s_1, s_2, \cdots, s_n\}$. $H(\cdot,\cdot)$ has the following properties: (1) Quasi-commutativity: $H(H(k, h_1), h_2) = H(H(k, h_2), h_1)$. (2) Absorbency: $H(H(k, h_i), h_i) = k \odot \alpha(h(h_i)) = H(k, h_i)$. (3) An item h_i within the accumulated value A can be verified by $H(A, h_i) = A \odot \alpha(h(h_i)) = A$.

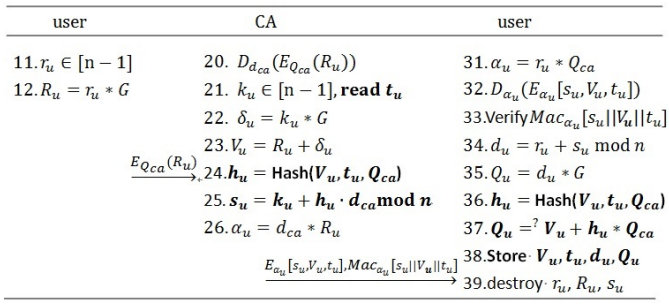


Fig. 3. The Proposed Secure and Efficient Implicit Certificate (SEIC) Scheme.

IV. The Proposed Scheme

In the section, we propose a secure and efficient implicit certificate (SEIC) scheme to overcome the P-HIP's weaknesses in Section II-A. In order to make the signature s_u prevent the forgery attack, the SEIC scheme constructs a secure digital signature algorithm by hashing the public key validation data V_u, the time-stamp t_u and the CA's public key Q_{ca}. The scheme also presents a table-and-optimality-based technique that makes the fixed-point multiplication in [4] more computationally efficient. Then, a certificate revocation mechanism is proposed. Finally, a formal proof for the security of SEIC is provided.

A. Proposed SEIC

The user sends a request R_u to the CA via a secure way (i.e.,public key encryption), by choosing a random integer $r_u \in [n-1]$ and then calculating $R_u = r_u * G$. Upon receiving the request, the CA obtains R_u. Then, by picking a random integer $k_u \in [n-1]$ and computing $V_u = R_u + k_u * G$ and s_u=$k_u + h_u \times d_{ca} \bmod n$, the CA issues a public key construction data s_u and a unique public-key authentication data V_u to the user, where h_u=$Hash(V_u, t_u, Q_{ca})$, t_u is the current time-stamp of the CA. Before sending (s_u, V_u, t_u) to the user, the CA uses the shared session key, $\alpha_u = d_{ca} * R_u$, to encrypt them and compute their Message Authentication Code (MAC). Upon receiving the messages from CA, the user computes α_u to decrypt $E_{\alpha_u}[s_u, V_u, t_u]$, and verifies $MAC_{\alpha_u}[s_u||V_u||t_u]$. Then, the user generates a private key d_u = r_u+$s_u \bmod n$ and public key $Q_u = d_u$*G. The user validates the result using the equality Q_u=V_u+$Hash(V_u, t_u, Q_{ca})$*Q_{ca}. The details of the proposed implicit certificate protocol are shown in Fig. 3.

The proposed SEIC can prevent the forged attacks in Section II-A. Since $Hash(\cdot)$ is one-way and collision-resistant, it is hard to compute a pre-image of a given value. That is, given a randomly chosen $h_u = Hash(V_u, t_u, Q_{ca})$, it is computationally infeasible to find a tuple $(\overline{V}_u, t_u, Q_{\overline{ca}})$ such that $Hash(\overline{V}_u, t_u, Q_{\overline{ca}}) = h_u$. Thus, based on (V_u, t_u, Q_{ca}), it is hard to forge an $\overline{V}_u$ satisfying Q_u=$\overline{V}_u + h_u * Q_{\overline{ca}}$. In other word, in the proposed SEIC, it is computationally infeasible for a malicious user holding the implicit certificate issued by the CA to forge an implicit certificate issued by another CA.

B. Performance Considerations

Assume that all the users know the system parameter pas=$\{\mathcal{G}, G, n, Hash(\cdot)\}$. The proposed SEIC can be very efficient since it allows a certain amount of precomputation.

1) Precomputation for $h_u * Q_{ca}$: Assume that b_m (bits) is the memory size used to the precomputation for $h_u * Q_{ca}$. Notice that the binary length of the output of $Hash$ is $|n|$, and the elliptic curve is on the finite field $\mathbb{F}_q$. We observe that the computational efficiency of $\hat{h} * Q_{ca}$ is significantly improved when $\hat{h}$ is restricted to a sufficiently small range. Note that the CA's public key Q_{ca} is commonly a fixed point for each user. Therefore, $h_u * Q_{ca}$ is amenable to optimization methods typical of fixed-point EC multiplications [4]. For a larger integer $h_u \in \{0,1\}^{|n|}$, we select a suitable base B and obtain its expansion (1) on the base B, so that each term $(c_l \cdot B^l) * Q_{ca}$ can be calculated efficiently.

$$h_u = c_\kappa \cdot B^\kappa + c_{\kappa-1} \cdot B^{\kappa-1} + \cdots + c_1 \cdot B + c_0. \quad (1)$$

Here, $0 \le c_l < B$, $l = 0, 1, \cdots, \kappa$, and $\kappa = \lceil |n|/(log_2 B)) \rceil$ is the number of substrings of length $log_2 B$ in h_u.

Given b_m, q and n, we design a table-and-optimality-based technique: how to choose an optimal base B such that the operation $h * Q_{ca}$ is accelerated. The specific operations are as follows:

(1) Define allowed values $AV = \{B = 2^\theta : 2|q|\lceil |n|/\theta \rceil 2^\theta < b_M\}$. This is because that there are $\kappa \cdot B$ intermediate results $(c_l \cdot B^l) * Q_{ca}$ to be stored, and $x, y \in \mathbb{F}_q$ for a point Q_{ca}=(x, y).

(2) It is recommended to select the largest $B = 2^\theta$ in AV. We notice that $\kappa = \lceil |n|/\theta \rceil$ decreases as $B = 2^\theta$ increases.

(3) By pre-computing $\hat{T}[l][c_l] = (c_l \cdot B^l) * Q_{ca}$ $(0 \le c_l < B, 0 \le l < \kappa)$and then storing them in the memory of the device, the h_u*Q_{ca} operation can be implemented via table look-ups as follows:

$$\begin{aligned} & h_u * Q_{ca} \\ &= (c_{\kappa-1} \cdot B^{\kappa-1} + \cdots + c_1 \cdot B + c_0) * Q_{ca} \\ &= (c_{\kappa-1} \cdot B^{\kappa-1}) * Q_{ca} + \cdots + (c_1 \cdot B^1) * Q_{ca} + c_0 * Q_{ca} \\ &= \hat{T}[\kappa-1][c_{\kappa-1}] + \cdots + \hat{T}[1][c_1] + \hat{T}[0][c_0]. \end{aligned}$$

This means that h_u*Q_{ca} operation can be attained through $(\kappa - 1)$ point additions.

For example, assuming that $|n|$=256, $|q|$=512, and b_m=512KBs=4194304 bits (IoT devices have a few megabytes of memory (8–32 KB of RAM and 48–512 KB of ROM), e.g., eZ1-Mote [12] has 32 KB of RAM and 512 KB of ROM). In the case, the allowed values is $AV = \{8, 16, 32, 64\}$. We choose B=64, and then κ=$\lceil 256/(log_2 64) \rceil$=43. Ignoring the (usually small) cost of table look-ups, this approach would take only 42 point additions. The size of the memory block storing the intermediate results is $43 \times 64 \times (512 + 512)$ bits =344 KBs.

In comparison, the method in SIMPL [4] using B=16 would require 63 point additions and 128 KBs memory. Thus,

the table-and-optimality-based technique is expected to be about 1.5 times faster than the method [4].

2) Implicit Certificate Revocation: Let $HIP^{(\sqcup)}$ be the set of revoked implicit certificates in time slot $\sqcup$. Based on the NOWA H, the revocation manager (RM) compute a NOWA value in $\mathbb{Z}_p$ by accumulating the hash values of implicit certificates in $HIP^{(\sqcup)}$. Then, the RM distributes the value in $\mathbb{Z}_p$ to all users in advance. Keeping just one $\mathbb{Z}_p$-symbol for the revocation verification reduces the storage and communication costs of each user. Specifically, when a user u requests to revoke her/his implicit certificate, the RM can revoke the implicit certificate as follows.

(1) The user u sends the implicit certificate (V_u, t_u, d_u, Q_u) and the CA's ID to the RM. The RM first determines the implicit certificate to be unexpired and correct by using the steps 36-37 in Fig. 3.

(2) The RM revokes the unexpired and correct implicit certificate by updating the previous $A_{\sqcup}$ with the new $A_{\sqcup'}$ for the time epoch $\sqcup'$, where $A_{\sqcup'} = H(A_{\sqcup}, \text{HIT}_u)$, and $\text{HIT}_u = Hash(Q_u)$. The RM then sends the value $A_{\sqcup'}$ to all users via the block chain or a tamper-proof electronic bulletin board.

(3) Each verifier downloads timely the new $A_{\sqcup'}$. The verifier then checks if $H(A_{\sqcup'}, \text{HIT}_u) \neq A_{\sqcup'}$ for the valid public key Q_u, where $\text{HIT}_u = Hash(Q_u)$. If the inequality holds, (V_u, t_u, Q_u) is valid; otherwise, it has been revoked.

C. Security Analysis

Under the assumption that the elliptic curve discrete logarithm problem (ECDLP) is hard on $\mathcal{G}$, we provide the security proof of SEIC as follows.

As shown in Fig. 3, the corresponding digital signature scheme DS=$(Gen, \mathcal{K}, \mathcal{S}, \mathcal{V})$ is defined as follows:

- $pas = \{\mathcal{G}, G, n, Hash(\cdot), H\} \leftarrow Gen(1^{\kappa})$: On inputting the security parameter κ, the probabilistic algorithm Gen outputs an array of system parameters pas.
- $(d_{ca}, Q_{ca}) \leftarrow \mathcal{K}(pas)$: On inputting the system parameters pas, the probabilistic algorithm $\mathcal{K}$ generates a pair of public and private keys (d_{ca}, Q_{ca}) for a CA.
- $(R_u, (V_u, t_u, s_u)) \leftarrow \mathcal{S}(d_{ca}, R_u)$: On inputting a private key d_{ca} and a message $R_u \in < G >$, the CA runs the probabilistic algorithm $\mathcal{S}$ to produce a signature σ = (V_u, t_u, s_u).
- $\{0, 1\} \leftarrow \mathcal{V}(Q_{ca}, R_u, \sigma)$: On inputting the CA's public key Q_{ca}, a message R_u and a signature σ, anyone can run the deterministic algorithm $\mathcal{V}$ to check whether σ is a valid signature. That is, σ is a valid signature if Q_u =$V_u + Hash(V_u, t_u, Q_{ca}) * Q_{ca}$, where Q_u = $R_u + s_u * G$.

The Lemma 1 in Appendix proves the unforgeability of DS in the proposed SEIC against adaptive chosen-message attacks.

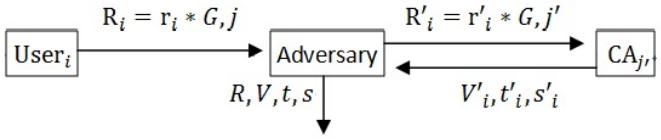


Fig. 4. Security Model for Implicit Certificate Adapted from [[7]].

Assume a scenario with n_{usr} legitimate users, denoted usr_i for $1 \leq i \leq n_{usr}$, and with n_{ca} CAs, denoted CA_j for $1 \leq j \leq n_{ca}$. Let $(R_i = r_i * G, j)$ denote usr_i's implicit certificate request for CA_j, and let (V_i, t_i, s_i) be the response sent by that CA. Also, let Q_i and d_i denote, respectively, the public and the private keys reconstructed by usr_i from CA_j's response, using that CA's public key Q_j. There are no restrictions on the number of credential requests that can be sent by usr_i to CA_j.

Definition 1. *A (τ', ϵ)-adversary $\mathcal{A}$ (of an implicit certificate scheme) is a probabilistic Turing machine that runs in time at most τ, interacting with legitimate users and CAs by performing each of the following operations any number of times:*

(1) *receive a request $(R_i = r_i * G, j)$ from usr_i for an implicit certificate from CA_j; and*

(2) *send a request $(R_{i'} = r_{i'} * G, j')$ to $CA_{j'}$, and receive response $(s_{i'}, V_{i'}, mac_{i'})$ from $CA_{j'}$.*

With probability at least ϵ, $\mathcal{A}$ outputs a triple (r, V, t, s) such that d=$r + s$ is the private key associated with the public key Q reconstructed from V and some Q_z (that is, $d * G$ = $V + Hash(V, t, Q_z) * Q_z$) such that either

(1) *[Forgery attack against CA_z]: (V, t, s) was never part of a response of CA_z for the request $(r * G, z)$; or*

(2) *[Key compromise against usr_i]: (V, t, s) was included in a response of CA_j to some request $(r * G, j)$ originally from usr_i, where $j \neq z$.*

A (τ', ϵ)-adversary is considered successful if ϵ is non-negligible for a polynomial time τ'.

In summary, as shown in Fig. 4, this model covers a scenario where the adversary $\mathcal{A}$ acts as proxy for requests from users and responses from CAs. Hence, $\mathcal{A}$ can: simply relay the request to the correct CA; modify the value of R_i =$r_i * G$ in the request; modify the user identifier i in the request, thus affecting the value of V in the credential; and/or forward the request to a different CA.

Under the security model of Definition 1(see Fig. 4) and the random oracle model, Theorem 3 in Appendix proves the security of the proposed SEIC.

V. Application to HIP in IoT

Public key validation can ensure the authenticity of an HIT in HIP. HIP is based on the Diffie-Hellman key exchange, using public key identifiers from a new host identity name-space for mutual peer authentication. The device uses a 128-bits hash of the public key as HIT.

An important challenge in HIP environments is to build a privacy-preserving HIP where authorized devices cannot be tracked, either by eavesdroppers or by the system itself [3]. One common approach for this issue is to provide a IoT device with multiple short-lived public keys. Then, IoT devices can avoid tracking by changing the public keys employed to sign its messages while it move from one network to another one. Hence, messages broadcast from different locations and using distinct public keys cannot be easily linked to any given IoT device. However, the total number of public keys valid simultaneously should be limited [4]. Among the existing solutions, the Secure Credential Management System (SCMS)[13] is one of the most relevant. The approach in SCMS combines privacy and scalability in the so-called butterfly key expansion process. Essentially, this process can issue multiple implicit certificates with a single request from a user. Furthermore, in the proposed SEIC, it reduces the amount of data exchanged and also the number of operations performed by the user.

In this section, we improve SEIC via the butterfly key expansion process, and then construct an improved P-HIP, We also formally analyze the privacy of the improved P-HIP.

A. Performance Improvement for SEIC

The implicit certificate issuance and revocation in the improved SEIC involves mainly four entities: User, Registration Authority (RA), CA and RM. Assume that there is no CA-RA collusion. They are respectively responsible for the following operations:

- User: the entity that requests credentials from a registration authority (RA). For better efficiency, each request leads to the provisioning of a batch containing β implicit certificates.
- RA: the entity that creates β implicit certificate requests to the CA from a single request of a user (called butterfly key expansion process). Those requests are individually forwarded to the CA, in such a manner that requests associated to different users are shuffled together.
- Credential Authority (CA): responsible for issuing credentials upon the requests by the RA. The credentials are then individually signed and encrypted by the CA before being sent back to the RA, from which they are delivered to the requesting user.
- Revocation Manager (RM): the entity that identifies the implicit certificates of users/devices and, whenever necessary, revokes them by accumulating their hash values to generate a value $A_{\sqcup}$ in $\mathbb{Z}_p$ (see Section IV-B2). Then, the RM distributes timely $A_{\sqcup}$ to all users in advance.

In the improved P-HIP, Fig. 5 presents the message exchange in the implicit certificate issuance phase, where the revocation operations are the same as that of the SEIC in Section IV-B2). All communications are made via secure ways, using standard protocols (e.g., Transport Layer Security-TLS) or public key encryption. The user first sends (R_u, f) to the RA. In response to the request of the user, the RA expands the point R_u into β points $\hat{R}_u^{(l)} = R_u + f(l) * G$. Note that f is shared only between the user and the RA. The RA then sends each individual $\hat{R}_u^{(l)}$ to the CA, while shuffling together requests from different batches to ensure their unlinkability.

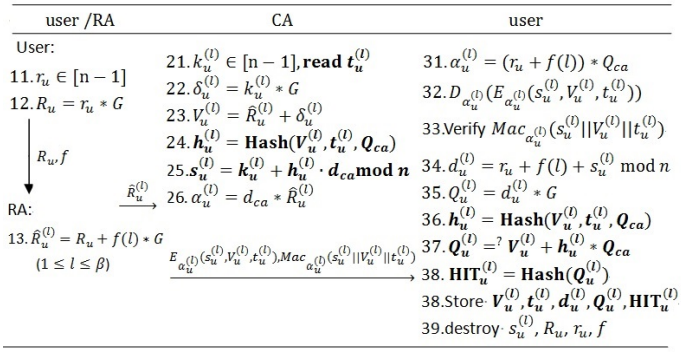


Fig. 5. The Implicit Certificate Issuance Phase in the Improved P-HIP.

The CA, in turn, randomizes $\hat{R}_u^{(l)}$ by picking randomly an integer $k_u^{(l)}$ and adding $k_u^{(l)} * G$ to it. The randomized point is used as the butterfly public key validation data $V_u^{(l)}$. Then, according to the procedure in Sections IV-A, the CA generates the hash value for $(V_u^{(l)}, t_u^{(l)}, Q_{ca})$, and outputs its signature $s_u^{(l)}$. The resulting credential is encrypted and verified with the Diffie-Hellman-key $\alpha_u^{(l)} = d_{ca} * \hat{R}_u^{(l)}$ and sent back to the RA. The RA, unable to decrypt the CA's response $\overline{pkg}$, forwards it back to the requesting user, in batch.

Finally, the user computes $\alpha_u^{(l)}$ to decrypt $\overline{pkg}$. It then verifies that the retrieved credential is indeed valid via $Mac_{\alpha_u^{(l)}}[s_u^{(l)} || V_u^{(l)}]$ and $Q_u^{(l)} = V_u^{(l)} + h_u^{(l)} Q_{ca}^{(l)}$, aiming to ensure there is no Man-in-the-Middle attack by the RA. If the verification is successful, the obtained keys, $(V_u^{(l)}, t_u^{(l)}, d_u^{(l)}, Q_u^{(l)})$, can be used for signing messages.

B. Performance Improvement for P-HIP

1) Host Identity and Host Identity Tag: Suppose that $u \in \{i, r\}$ is an authorized user as an initiator i or a responder r. Then, β implicit certificates $\{V_u^{(l)}, t_u^{(l)}, d_u^{(l)}, Q_u^{(l)}\}$ are first obtained via the process in Section V-A, l=1, $\cdots$, β. For each implicit certificate $\{V_u^{(l)}, t_u^{(l)}, d_u^{(l)}, Q_u^{(l)}\}$, its host identity is the public key $Q_u^{(l)}$. The corresponding HIT can be computed by the user u as $\text{HIP}_u^{(l)} = Hash(Q_u^{(l)})$. Hence, the mobile user u can use a new host identity $Q_u^{(l)}$ and its $\text{HIP}_u^{(l)}$ to avoid identity tracking when she/he moves from one network to another network.

2) Host Identity Validation: Here, we present the procedure to validate host identifiers, such as public keys and HITs. Without loss of generality, assume that a prover is the user u who holds an implicit certificate $\{V_u^{(l)}, t_u^{(l)}, d_u^{(l)}, Q_u^{(l)}\}$ and $\text{HIP}_u^{(l)}$. Now, the prover provides its host identity $Q_u^{(l)}$, $\text{HIP}_u^{(l)}$, the public-key authentication data $(V_u^{(l)}, t_u^{(l)})$ and the CA's identity to a verifier. The verifier validates u's implicit certificate by running Algorithm 1 in Fig. 6 as $PubKeyValid(pas, Q_{ca}, A_{\sqcup}, (V_u, t_u, Q_u, HIP_u))$.

Algorithm 1 : Public kry validation
$\{0,1\} \leftarrow PubKeyValid(pas, Q_{ca}, A_{\sqcup}, (V_u, t_u, Q_u, HIP_u))$: It takes pas,Q_{ca}, $A_{\sqcup}$ and (V_u, t_u, Q_u, HIP_u) as the input, and outputs 1 if (Q_u, HIP_u) is authentic or 0 if either the public-key or HIT is forged.
(1) t_u is unexpired.
(2) Q_u is unrevoked by computing $HIT_u^{(l)} = Hash(Q_u)$ and then check if $H(A_{\sqcup}, HIT_u^{(l)}) \neq A_{\sqcup}$.
(3) Q_u is correct and issued by CA if $Q_u^{(l)}$=$V_u^{(l)} + h_u^{(l)} * Q_{ca}$ holds.
(4) If the above checks are true, it returns 1; otherwise, it return 0.

Algorithm 2 : Keying material generation
$null$ or $((d_{pu}, Q_{pu}), V_{pu}) \leftarrow KeyMatGen(HIT_u, d_u, PuzSol)$: It takes the initiator/responder host identifier HIT_u, her/his private key d_u and puzzle/solution $PuzSol$ as the input, and outputs $null$ or a message $((d_{pu}, Q_{pu}), V_{pu})$ as follows:
If HIT_u equals $null$ or HIT_u is revoked or $PuzSol$ equals $null$, it returns $null$; otherwise, it returns $((d_{pu}, Q_{pu}), V_{pu})$ by the following operations.
(1) Choose a random integer $r_{pu} \in [n-1]$ and generate a random point V_{pu} = $r_{pu} * G$.
(2) Compute the hash value h_{pu} = $Hash(HIT_u || PuzSol || V_{pu})$.
(3) Use the private key d_u to generate V_{pu}'s validation data Q_{pu} = $d_{pu} * G$, where d_{pu}=$h_{pu} \times d_u + r_{pu} \mod n$.

Algorithm 3 : Keying material validation
$null$ or $\{0,1\} \leftarrow KeyMatValid(HIT_u, Q_u, PuzSol, Q_{pu}, V_{pu})$: It takes the initiator/responder host identifier HIT_u, her/his public key Q_u, puzzle/solution $PuzSol$, the random point V_{pu}, and the validation data Q_{pu} as the input, and outputs $null$ or 1 if V_{pu} is authentic or 0 if the random point is forged. Specifically, if HIT_u equals $null$ or $PuzSol$ equals $null$, it returns $null$; otherwise, it returns $\{0,1\}$ by the following operations.
(1) Compute the hash value h_{pu} = $Hash(HIT_u || PuzSol || V_{pu})$.
(2) Generate the point Q'_{pu} = $h_{pu} * Q_u + V_{pu}$.
(3) If Q'_{pu} = Q_{pu}, it returns 1; otherwise, it returns 0.

Fig. 6. The Algorithms in the Improved HIP

The correctness of the host identity $Q_u^{(l)}$ can be seen as follows. $Q_u^{(l)} = d_u^{(l)} * G = (r_u + f(l) + s_u^{(l)}) * G = (r_u + f(l) + k_u^{(l)} + h_u^{(l)} \cdot d_{ca}) * G = \hat{R}_u^{(l)} + \delta^{(l)} + (h_u^{(l)} \cdot d_{ca}) * G = V_u^{(l)} + h_u^{(l)} * Q_{ca}$.

3) Mutual Authentication: Let the CA issue $(V_i^{(l)}, t_i^{(l)}, d_i^{(l)}, Q_i^{(l)})$ and $(V_r^{(\iota)}, t_r^{(\iota)}, d_r^{(\iota)}, Q_r^{(\iota)})$ to the initiator and responder, respectively. Assume that CA's public key Q_{ca} and RM's authentication information $A_{\sqcup}$ are correctly sent to all users in advance. The initiator and responder compute $\text{HIT}_i^{(l)}$ and $\text{HIT}_r^{(\iota)}$ as Section V-B1), respectively. For the improved P-HIP, the mutual authentication procedure is shown in Fig. 7.

In the improved P-HIP,the operations in the first step and the last step are the same to the common HIP. We will omit the statement of these two steps, and will describe the other intermediate steps and operations in detail.

Upon receiving the message I_1, the responder then creates a message R_1 that contains $\text{HIT}_i^{(l)}$, $\text{HIT}_r^{(\iota)}$, a puzzle, its keying material (Q_{pr}, V_{pr}), its host identity $Q_r^{(\iota)}$ and the corresponding validation data $(V_r^{(\iota)}, t_r^{(\iota)})$. The keying material is generated by the responder running the Algorithm 2 in Fig. 6 as $((d_{pr}, Q_{pr}), V_{pr}) \leftarrow KeyMaterialGen(HIT_r^{(\iota)}, d_r^{(\iota)}, puzzle)$.

After receiving the message R_1, the initiator first validates

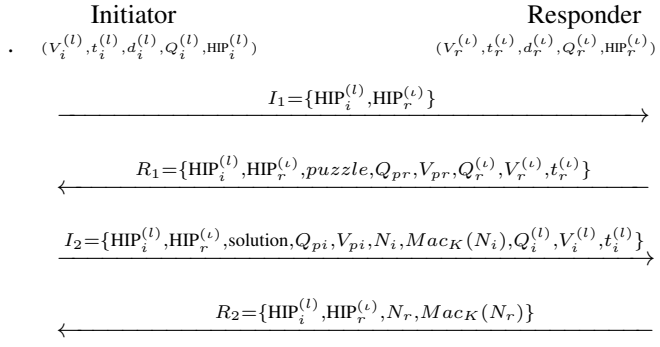


Fig. 7. The Mutual Authentication in the Improved P-HIP.

the host identity $Q_r^{(\iota)}$ and $\text{HIT}_r^{(\iota)}$ by using the Algorithm 1 in Fig. 6 to ensure that the keying material (Q_{pr}, V_{pr}) is generated by an authorized user. Next, the initiator solves the puzzle to get its solution, and then verifies the keying material (Q_{pr}, V_{pr}) in R_1 by running the Algorithm 3 in Fig. 6, this is, $\{0,1\} \leftarrow KeyMatValid(HIT_r^{(\iota)}, Q_r^{(\iota)}, puzzle, Q_{pr}, V_{pr})$. The initiator generates the keying material by using the Algorithm 2 in Fig. 6 as $((d_{pi}, Q_{pi}), V_{pi}) \leftarrow KeyMatGen(HIT_i^{(l)}, d_i^{(l)}, solution)$. Finally, the initiator computes a session key K=$d_{pi} * Q_{pr}$, and sends the message I_2 to the responder. The message I_2 includes the elements as shown in Fig. 7. If one of the above checks returns false, the initiator exits.

Similar to the initiator's approaches, the responder verifies the authenticity of the host identity $Q_r^{(\iota)}$ and $\text{HIT}_r^{(\iota)}$, validates the keying material (Q_{ir}, V_{ir}) in I_2, and authenticates the initiator. Next, the responder computes the session key using d_{pr} and Q_{ir} as $K = d_{pr} * Q_{ir}$, and then validates the $MAC_K(N_i)$.

Finally, at the last step, the responder sends R_2 and the initiator validates $MAC_K(N_r)$ to ensures that a shared session key K is created successfully. After this point, the communications between them are encrypted using K.

4) User Privacy: In the improved P-HIP, a user may utilize a remaining unused implicit certificate to generate the new host identity while she/he moves to a different network or wants to update its host identity. This way provides a secure interact between different parties without compromising user privacy. Let us suppose, an adversary is provided with the implicit certificates, $\{(V_u^{(l)}, t_u^{(l)}, d_u^{(l)}, Q_u^{(l)})\}_{l=1}^{\varpi}$, of a user u that were used in ϖ number of networks. The goal of the adversary is to infer that the implicit certificates belong the same user. In Section VI-B, we utilize the formal privacy definition [15], to prove that the improved P-HIP gives rigorous, rather than ad-hoc or intuition-based privacy guarantees.

VI. The Security Analysis of the Improved P-HIP

In this section, we analyze the security of the improved SEIC and the privacy of improved P-HIP.

A. The Security Analysis for Improved SEIC

Here, we show that the improved SEIC in Section V-A is secure. Under the attack model in Definition 1, Theorem 3 shows that there is no adversary $\mathcal{A}$ that is successful against the proposed SEIC. Form Lemma 1, we know that SEIC's signature scheme DS is secure against adaptive chosen-message attacks. The improved SCMS in Section is the butterfly key expansion of the proposed SEIC. The attacks in Definition 1 does not invalidate SEIC's security claims for at least three reasons[4]. The first is that in the improved SEIC, one of security assumptions is that there is no CA-RA collusion. Next, SCMS recommends using the ECDSA-signature algorithm [13], for which SEIC's signature scheme DS is a secure ECDSA-signature against the attacks in Definition 1. Finally, the latest version of SCMS already suggests the countermeasure hereby proposed [14], that is, the signer's certificate information is included in the hash computation. Therefore, the improved SEIC remains secure against the forgery in Definition 1.

B. The Formal Privacy Analysis of the Improved P-HIP

We first define the privacy model, and then formally analyze the privacy of the improved P-HIP.

1) Privacy Model: We now consider Ouafi and Phan's privacy model [15]. In this model, attacker $\mathcal{A}$ can eavesdrop on all the channels between two users, and he/she can also perform any active or passive attacks. In this regard, $\mathcal{A}$ needs to model the following queries in polynomial time:

Execute$(\mathcal{P},\mathcal{U},s)$: This query represents the passive attacks. In this context, the attacker can eavesdrop all the transmitted messages between the user $\mathcal{U}$ and a party $\mathcal{P} \in \{CA, RA, \mathcal{V}\}$ in the s-th session, where the user $\mathcal{V}$ satisfies $\mathcal{V} \neq \mathcal{U}$. Consequently, the attacker obtains all the exchanged data between $\mathcal{U}$ and $\mathcal{P}$.

Send$(\mathcal{U},\mathcal{V},m,s)$: This query models the active attacks in the system. In this query, attacker $\mathcal{A}$ has the permission to impersonate a user $\mathcal{U}$ in the s-th session, and forwards a message m to another user $\mathcal{V}$. Besides, the attacker has the permission to block the exchanged message m between $\mathcal{U}$ and $\mathcal{V}$.

Query$(\mathcal{U},m_1,m_2)$: This query models the adversary's ability to investigate a user. For this, $\mathcal{A}$ sends m_1 to $\mathcal{U}$ and receives m_2 from $\mathcal{U}$.

Corrupt$(\mathcal{U},K)$: In this query, the attacker $\mathcal{A}$ has the permission to access secret information K stored in the user $\mathcal{U}$'s memory.

Test$(\mathcal{U}_0,\mathcal{U}_1,s)$: This query is the only query that does not correspond to any of $\mathcal{A}$'s abilities or any real-world event. This query allows to define the indistinguishability-based notion of untraceable privacy.

If the party has accepted and is being asked a Test query, then depending on a randomly chosen bit $b \in \{0,1\}$, $\mathcal{A}$ is given $\mathcal{U}_b$ from the set $\{\mathcal{U}_0,\mathcal{U}_1\}$. Informally, $\mathcal{A}$ succeeds if it can guess the bit b. In order for the notion to be meaningful, a Test session must be fresh in the sense of Definition 3.

Definition 2 (Partnership and Session Completion)**.** *An initiator instance i and a responder instance r are partners if, and only if, both have output Accept(i) and Accept(r), respectively, signifying the completion of the protocol session.*

Definition 3 (Freshness)**.** *A party instance is fresh at the end of execution if, and only if (1) it has output Accept with or without a partner instance and (2) both the instance and its partner instance (if such a partner exists) have not been sent a Corrupt query.*

Definition 4 (Indistinguishable Privacy (INDPriv))**.** *It is defined using the game $\mathcal{G}$ played between a malicious adversary $\mathcal{A}$ and a collection of initiators and responders and RAs and CA instances. $\mathcal{A}$ runs the game $\mathcal{G}$ whose setting is as follows.*

- *Learning phase: $\mathcal{A}$ is able to send any Execute, Send, Query, and Corrupt queries and interact with the RA, the CA and users $\mathcal{U}_0$, $\mathcal{U}_1$ that are chosen randomly.*
- *Challenge phase: The attacker selects two users $\mathcal{U}_0$ and $\mathcal{U}_1$, and forwards a Test query $(\mathcal{U}_0,\mathcal{U}_1,s)$ to challenger $\mathcal{C}$. After that, $\mathcal{C}$ randomly selects $b \in \{0,1\}$ and the attacker determines a user $\mathcal{U}_b \in \{\mathcal{U}_0,\mathcal{U}_1\}$ using Execute, Send and Query queries.*
- *Guess phase: The attacker $\mathcal{A}$ finishes the game $\mathcal{G}$ and outputs a bit $\hat{b} \in \{0,1\}$ as guess of b. The success of attacker $\mathcal{A}$ in the game $\mathcal{G}$ and consequently breaking the security of INDPriv is quantified via $\mathcal{A}$'s advantage in recognizing whether attacker $\mathcal{A}$ received $\mathcal{U}_0$ or $\mathcal{U}_1$, and is denoted by $Adv_{\mathcal{A}}^{INDPriv}(k)$ = $|Pr[\hat{b}=b]-1/2|$, where k is a security parameter.*

Theorem 1. *The improved P-HIP satisfies indistinguishable privacy.*

Proof 1. *In the improved P-HIP, after a successful authentication, the user $\mathcal{U}_0$ update its secret key $d_{\mathcal{U}_0}$. Besides, the host identities $HIP_{\mathcal{U}_0}$ change in each session. Therefore, it will be difficult for an adversary to perform any traceability attack by performing the following phases:*

- *Learning phase: In the μ-th authentication instance, the adversary $\mathcal{A}$ is able to send any Execute$(CA,\mathcal{U}_0,\mu)$ queries and obtains the public key $Q_{\mathcal{U}_0}^{(\mu)}$ and the host identity $HIP_{\mathcal{U}_0}^{(\mu)}$ such that $HIP_{\mathcal{U}_0}^{(\mu)}$=$Hash(Q_{\mathcal{U}_0}^{(\mu)})$ holds.*
- *Challenge phase: The adversary $\mathcal{A}$ selects two fresh users $\mathcal{U}_0$, $\mathcal{U}_1$ and forwards a Test query $(\mathcal{U}_0,\mathcal{U}_1,\mu+1)$ to the challenger $\mathcal{C}$. Next, according to the randomly chosen bit $b \in \{0,1\}$, $\mathcal{A}$ is given a user $\mathcal{U}_b \in \{\mathcal{U}_0,\mathcal{U}_1\}$. After that the adversary $\mathcal{A}$ sends a query Execute$(CA,\mathcal{U}_b,\mu+1)$ and obtains the public key $Q_{\mathcal{U}_b}^{(\mu+1)}$ and the host identity $HIP_{\mathcal{U}_b}^{(\mu+1)}$, where $HIP_{\mathcal{U}_b}^{(\mu+1)}$=$Hash(Q_{\mathcal{U}_b}^{(\mu+1)})$.*
- *Guess phase: In the Learning phase the user $\mathcal{U}_0$ updates its secret $d_{\mathcal{U}_0}$, therefore for the two subsequent sessions μ and $\mu+1$, the public keys $Q_{\mathcal{U}_0}^{(\mu)}$=$d_{\mathcal{U}_0}^{(\mu)} * G$ and $Q_{\mathcal{U}_b}^{(\mu+1)} = d_{\mathcal{U}_b}^{(\mu+1)} * G$ are calculated as follows: $d_{\mathcal{U}_0}^{(\mu)} = [(r_{\mathcal{U}_0} + f(\mu)) + k_{\mathcal{U}_0}^{(\mu)} + Hash((r_{\mathcal{U}_0} + f(\mu)) * G, t_{\mathcal{U}_0}^{(\mu)}, Q_{ca})d_{ca}] \mod n$, $d_{\mathcal{U}_b}^{(\mu+1)} = [(r_{\mathcal{U}_b} + f(\mu+1)) + k_{\mathcal{U}_b}^{(\mu+1)} + Hash((r_{\mathcal{U}_b} + f(\mu+1)) *$*

$G, t_{\mathcal{U}_b}^{(\mu+1)}, Q_{ca})d_{ca}] \mod n$. *Note that* $r_{\mathcal{U}_0}$, $k_{\mathcal{U}_0}^{(\mu)}$, $r_{\mathcal{U}_b}$ *and* $k_{\mathcal{U}_b}^{(\mu+1)}$ *are the random numbers. Assume that the hash function* $Hash$ *is truly random, mapping each data item independently and uniformly to the range* $\{0,1\}^r$, *that is,* $Pr[Hash(x) = Hash(x')] = \frac{1}{2^r}$ *where* $x \neq x'$. *Since* $t_{\mathcal{U}_0}^{(\mu)} \neq t_{\mathcal{U}_b}^{(\mu+1)}$, *therefore* $d_{\mathcal{U}_b}^{(\mu+1)} = d_{\mathcal{U}_b}^{(\mu+1)}$ *with the probability less than* $\frac{1}{2^{r-1}}$. *In other word,* $Q_{\mathcal{U}_0}^{(\mu)} = Q_{\mathcal{U}_b}^{(\mu+1)}$ *holds with the probability less than* $\frac{1}{2^{r-1}}$. *Again,* $HIP_{\mathcal{U}_0}^{(\mu)}$=$Hash(Q_{\mathcal{U}_0}^{(\mu)})$ *and* $HIP_{\mathcal{U}_b}^{(\mu+1)}$=$Hash(Q_{\mathcal{U}_b}^{(\mu+1)})$. *Hence, the adversary needs make a random guess for* $HIP_{\mathcal{U}_b}^{(\mu+1)}$. *In this context, the advantage of the adversary recognizing* $\mathcal{U}_0$ *or* $\mathcal{U}_1$, *can be denoted* $Adv_{\mathcal{A}}^{INDPriv}(k) = |Pr[\hat{b} = b] - 1/2| < \epsilon$, *where* $\epsilon = \frac{1}{2^{r-1}}$ *is negligible when* r *is large enough.*

VII. Performance Analysis and Comparison

In this section, based on the improved scheme in Section V, we compare it with other similar solutions in the literature in terms of the desired security properties, computation cost and communication cost.

A. Performance Comparison

A comparison of the security properties among the improved scheme with other implicit certificate schemes [3, 4, 13] is given in Table II. The improved scheme in Section V is secure against the forger and credential misbinding attacks. The signature algorithm's input (see step 24 Fig. 3 and 5) includes both $(V_u^{(l)}, t_u^{(l)})$ and the signer's public key Q_{ca}. In the P-HIP scenario, enforcing this technique when signing public-key authentication data can avoid forgery attacks that builds upon the properties of butterfly keys. Under the attack model in Definition and Ouafi and Phan's privacy model, the improved scheme provides the rigorous security proof in Section IV-C and VI-A and the formal privacy analysis in Section VI-B, respectively. In addition, the improved scheme achieves the revocation verification of a public key by performing one NOWA operation in Section IV-B2. However, the schemes [3, 4, 13] focused on the informal analysis of user privacy, and did not consider the revocation of unexpired implicit certificates. Note that the formal security proof did not provided in [3].

TABLE II. Performance Comparison based on Security Properties with Respect to Implicit Certificate Schemes

Scheme	SP1	SP2	SP3	SP4	SP5
ECQV [13]	No	formal	informal	No	-
SIMPL [4]	Yes	formal	informal	No	SCMS
P-HIP [3]	No	informal	informal	No	HIP
SEIC	Yes	formal	No	Yes	HIP
improved P-HIP	Yes	formal	formal	Yes	SCMS, HIP
SP1: preventing the credential misbinding attacks and a forgery attack					
SP2: security proof; SP3: user privacy proof;					
SP4:credential revocation; SP5: Compatibility					

B. Effectiveness Analysis

We evaluate the effectiveness of the improved scheme in terms of the computation and communication costs.

1) Experimental Results: To show the effectiveness of the improved scheme with respect to the existing implicit certificate schemes, we conduct simulations of the cryptographic operations used by various schemes on an Intel(R) Core(TM) i7-8550U CPU@1.80 GHz laptop computer with 8.00 GB memory and Windows10 using JDK1.8 (operating as the initiator or the responder as per the scheme). The simulations used the JPBC library jpbc-2.0.0 [16] to evaluate the execution time of different cryptographic operations.

We create an ECC self-signed X.509 certificates using the type A pairings on the curve $y^2 = x^3 + x$ over the finite field $\mathbb{F}_q$. SHA-256 is chosen as the cryptographic hash function $Hash$. In addition, we select SHA-512 for hashing h in NOWA H with a 128 bit output, where $N = 2^4$ is an upper bound to the number of accumulated items. When $N > 2^4$, we do this by selecting $\eta = \lceil N/(2^4) \rceil$ different SHA-512 as Remark 1 in [17]. For the function $Hash$ and the message authentication code (MAC), the SHA-256 is chosen as suggested. Furthermore, the leftmost 128 bits in the output of $Hash(Q_u)$ is taken as a HIT corresponding public key Q_u. With the above parameter settings, we consider the average value of over 100 trials for an operation $o \in \{Hash, H, a(\text{Point addition}), m(\text{Point multiplication}), e(\text{AES encryption}), d(\text{AES decryption})\}$. The results are as follows: T_{Hash} = 1.2828milliseconds(ms), T_H = 53.8039(ms), T_a = 1.3418(ms), T_m = 96.9339(ms), T_e = 13.1607(ms), and T_d = 3.7243(ms). In particular, the average time performing an addition or a multiplication of two numbers is 0.6626ms or 0.7615ms, which is negligible compared to other operations.

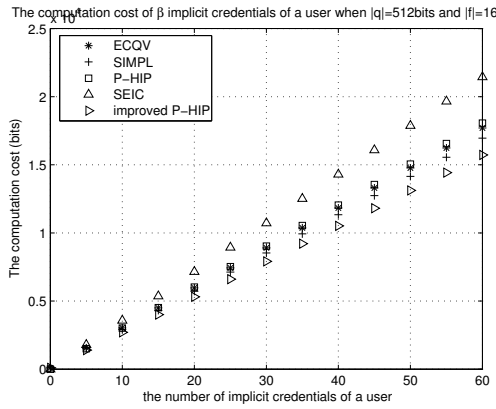
2) Implicit Certificate Issuance:

a) Computation Cost: Computation costs are the principal constraint for IoT users/devices, and we show a reduction in required computation in the improved P-HIP as compared to the existing schemes. Table III shows the computation cost of a user in different schemes. The improved P-HIP is similar to the approach discussed in [4], the key difference is that instead of 63 point addition, $h * Q_{ca}$ is computed by using 42 point addition. In both schemes, the user generates the request $(R_u = r_u * G, f)$, and then obtains an implicit certificate the β by computing $\alpha_u^{(l)}$, $d_u^{(l)} * G$ and $h_u^{(l)} * Q_{ca}$. It means that in the improved P-HIP, the computation cost of the user is $\beta(2\text{T}_m + 44\text{T}_a + T_d + 3\text{T}_{Hash}) + \text{T}_m$. In addition, the verifier performs only one H operation to check whether the unexpired implicit certificates is unrevoked. From Fig. 8 (a), it is evident that the computation cost of a user increases with the number of implicit certificate, but it grows relatively slowly in the improved scheme. In particular, the improved P-HIP makes a user has the smallest computation cost.

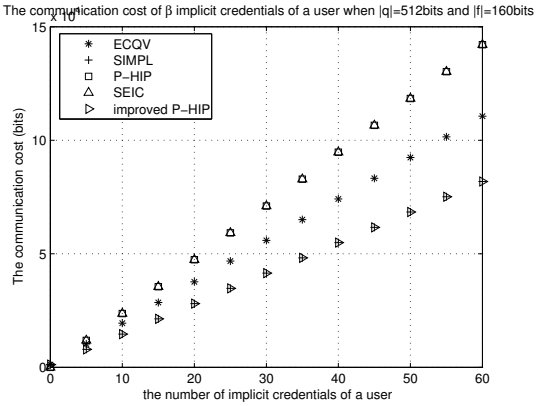
TABLE III. The Cost Comparison of Different Schemes in the Implicit Certificate Issuance Phase

Scheme	Computation cost	Communication cost
ECQV[13]	$\beta(3\text{T}_m + \text{T}_a + \text{T}_{Hash}) + \text{T}_m$	$\geq \beta(2\|q\| + 800) + 2\|q\| + \|f\|$
SIMPL[4]	$\beta(2\text{T}_m + 64\text{T}_a + \text{T}_{Hash}) + \text{T}_m$	$\geq \beta(2\|q\| + 320) + 2\|q\| + \|f\|$
P-HIP [3]	$\beta(3\text{T}_m + 2\text{T}_a + \text{T}_d + 2\text{T}_{Hash})$	$\beta(4\|q\| + 320)$
SEIC	$\beta(3\text{T}_m + 44\text{T}_a + \text{T}_d + 3\text{T}_{Hash})$	$\beta(4\|q\| + 320)$
Improved P-HIP	$\beta(2\text{T}_m + 44\text{T}_a + \text{T}_d + 3\text{T}_{Hash}) + \text{T}_m$	$\beta(2\|q\| + 320) + 2\|q\| + \|f\|$

b) Communication Cost: The advantage of the improved scheme is that the communication cost is low for IoT users/devices in the implicit certificate issuance. The communication cost comparison of these schemes [3, 4, 13] is



(a) The Computation Cost of β Implicit Certificates of a User



(b) the Communication Cost of β Implicit Certificates of a User

Fig. 8. The Computation and Communication Costs of β Implicit Certificates of a User in the Improved P-HIP

shown in Table III. To give a detailed quantitative analysis, we create a ECC self-signed X.509 certificates using the OpenSSL library [2], and choose n = 160 bits and q = 512 bits. The sizes of an identity and a time-stamp are recommended to be 20 bytes [14]. In the improved P-HIP, the communication cost at the user is as follows: The size of β responses from the CA $(c_u^{(l)}, Mac_u^{(l)})$ are $\beta(2|q|+n+320)$ bits, and the size of a request is $|R_u|+|f|$=2$|q|$+f bits. The total communication cost of a user is $(2\beta+2)|q|+320\beta+|f|$ bits. We notice that the size of $meta$ in [4, 13] is not less than 160 bits since $meta$ contains at least a time-stamp. The ECDSA-based signature outputs at least two numbers in $\mathbb{F}_q$. The length of $sig_u^{(l)}$ in ECQV[13] is $2|q|$. However, the size of $|sig_u^{(l)}|$ is $|q|$ since $sig_u^{(l)}$ is a number in $\mathbb{Z}_n^*$. Fig. 8 (a) and (b) shows that the improved scheme makes a user have both the smallest computation cost and the smallest communication cost.

3) Mutual Authentication: Implicit Credential Validation: We also evaluate the benefits of the improved P-HIP when users perform mutual authentication. The gains in this case originate from the following observations.

In Section V-B2), the cost of validating a public key and its HIT is T_H+ $\hat{T}_m$+ T_a+2T_{Hash}, where $h_u^{(l)} * Q_{ca}$ can be accelerated at cost $\hat{T}_m$ = 42T_a< T_m. In addition, the costs for keying material generation and keying material validation (see Algorithm 2 and 3 in Fig. 6) are 2T_m+ T_a+ T_{Hash}) and T_m+ T_a+ T_{Hash}, respectively. However, the generation/validation of a MAC requires one hash operation. Therefore, the total computational cost of an initiator or a responder is T_H +3T_m+ 40T_a+ 6 T_{Hash} $\approx$ 411.3416ms.

The improved HIP has stronger capabilities (such as public key revocation verification) than P-HIP [3], with an approximate computational cost and the same communication cost. In P-HIP, an initiator or a responder takes the total computational cost to be 4T_m+ 3T_a+ 6 T_{Hash} $\approx$ 399.4578ms. This value is close to the computational cost of the user in the improved P-HIP. On the other hand, the time for the device to perform a public key revocation verification operation is 53.8039(ms). Since $A_{\sqcup}$ is sent to all users in advance, both schemes have the same communication cost. The result gives a glimpse of SEIC's potential to speed up both signature verification and public-key authentication in HIP environments.

VIII. Conclusion

In this article, we propose SEIC that can improve the security of the P-HIP and the efficiency of EC point multiplications for IoT devices. For fix-point multiplication, the proposed method is about 1.5 times faster than the method in SIMPL scheme. At the same time, by making use of the butterfly key expansion process, we constructs an improved P-HIP by reducing the amount of data exchanged and the number of operations performed by user. Our evaluation shows that the improved P-HIP helps to achieve all the important security properties and ensures the user privacy with reasonable computation cost. However, one limitation is that for the solution to become a reality, the revocation verification information $A_{\sqcup}$ must be updated timely and sent to all users in advance, which can be costly. A potential future direction of this research is how to effectively implement the update and release for $A_{\sqcup}$. Note that the proposed scheme has minimized the revocation verification information because $A_{\sqcup}$ is just a $\mathbb{Z}_p$ symbol. Hence, for a IoT network with a small number of users, the update of $A_{\sqcup}$ can be effectively implemented. How to achieve the update of $A_{\sqcup}$ for all users in the large and medium-sized IoT networks is still a challenge problem.

Acknowledgment

This work is supported in part by National Key Research and Development Plan of China under Grants 2018YFB1003701, and in part by the 22nd batch of Teaching Reform Research Project of Jinan University under Grant 55611518.

References

[1] X. Shi, S. Shi, M. Wang, J. Kaunisto, C. Qian, On-device IoT Certificate Revocation Checking with Small Memory and Low Latency. CCS 2021: 1118-1134.

[2] J. Viega, M. Messier, P. Chandra, Network security with OpenSSL: Cryptography for secure communications, Sebastopol, CA, USA, O'Reilly Media, 2002.

[3] M. M. Hossain, R. Hasan. P-HIP: A lightweight and privacy-aware host identity protocol for Internet of Things. IEEE Internet Things J. 8(1): 555-571 (2021)

[4] P. S. L. M. Barreto, M. A. Simplício Jr., J. E. Ricardini, H. K. Patil, Schnorr-based implicit certification: Improving the security and efficiency of vehicular communications, IEEE Trans. Computers 70(3): 393-399 (2021).
[5] C.G. Günther, An identity-based key-exchange protocol, in: Advances in Cryptology-Eurocrypt'89, 1989, pp. 29-37.
[6] M. Girault, Self-certified public keys, In: Advances in Cryptology-Eurocrypt'91, 1991, pp. 490-497.
[7] D. Brown, R. Gallant, S. Vanstone, Provably secure implicit certificate schemes, in Proc. 5th Int. Conf. Financial Cryptography, 2002, pp. 156–165.
[8] M. Campagna, Sec 4: Elliptic curve qu-vanstone implicit certificate scheme (ECQV), Certicom Res., vol. 4, p. 32, 2013.
[9] P. Nikander, A. Gurtov, T. R. Henderson, Host identity protocol (HIP): connectivity, mobility, multi-homing, security, and privacy over IPv4 and IPv6 networks, IEEE Commun. Surv. Tutorials 12(2): 186-204 (2010)
[10] G. Montenegro, N. Kushalnagar, J. Hui, D. Culler, Transmission of IPV6 packets over IEEE 802.15.4 networks, IETF, RFC 4944, 2007.
[11] K. Nyberg, Fast accumulated hashing, in: Proc. 3rd Int. Workshop Fast Softw. Encryption, 1996, pp. 83–87.
[12] Zolertia, Z1 Mote IoT Device, 2016. [Online]. Available: http://zolertia.sourceforge.net/
[13] B. Brecht, D. Therriault, A.Weimerskirch, W. Whyte, V. Kumar, T. Hehn, R. Goudy, A security credential management system for V2X communications, IEEE Trans. Intell. Transp. Syst. 19(12): 3850-3871 (2018)
[14] NIST, FIPS 186–4: Digital Signature Standard (DSS), National Institute of Standards and Technology, Jul. 2013. [Online]. Available: http://nvlpubs. nist.gov/nistpubs/FIPS/NIST.FIPS.186–4.pdf.
[15] P. Gope, B. Sikdar, Lightweight and privacy-friendly spatial data aggregation for secure power supply and demand management in smart grids, IEEE Trans. Inf. Forensics Secur. 14(6): 1554-1566 (2019).
[16] PBC library, accessed: Apr. 16, 2017. [Online]. Available: http://crypto.standford.edu/pbc/.
[17] L. Lu, J.Lu, A lightweight verifiable secret sharing in Internet of Things, IJACSA 13(5):1028-1035 (2022).

APPENDIX

Lemma 1. *Assuming* (1) *ECDLP is hard for* $\mathcal{G}$ *and* (2) $Hash$ *is a random oracle, the Schnorr signature scheme* DS *in Fig. 3 is secure against adaptive chosen-message attacks.*

Proof 2. *Let* $Adv^{cma}_{DS,\mathcal{A}}(\kappa)$ *be the probability that* $\mathcal{A}$ *breaks the above Schnorr signature scheme* DS *and achieve a forgery. Assume that* $Adv^{cma}_{DS,\mathcal{A}}(\kappa)$ *is nonnegligible. We will construct an algorithm* $\mathcal{B}$ *which can solve the ECDLP in* $\mathcal{G}$*.*

Let G *be a generator of* $\mathcal{G}$*. Given a point* Q_{ca}=$d_{ca} * G \in <G>$ *as a challenge to* $\mathcal{B}$*, it aims to output such a value* $d_{ca} \in Z^*_q$*. The hash function* $Hash$ *behaves as a random oracle.*

$\mathcal{B}$ *starts* $\mathcal{A}$ *on input* 1^κ*. Let* $T_1(\kappa)$ *denote the bound of the number of CAs.* $\mathcal{B}$ *picks at random a* $i \in \{1, \cdots, T_1(\kappa)\}$*, guessing that* $\mathcal{A}$ *will succeed against the entity* i*.* $\mathcal{B}$ *runs* $\mathcal{K}(pas)$ *to generate for each entity its private/public pair, except* i*.* i *is given a public key* Q_{ca} *, while the corresponding private key* d_{ca} *is unknown to* $\mathcal{B}$*. A* i*'s signature on a message can be generated by querying the signing oracle* $\mathcal{S}(d_{ca}, \cdot)$*.*

$\mathcal{B}$ *can simulate the entity* i *to respond messages to via the following oracles:*

Hash Queries: *At any time,* $\mathcal{A}$ *can query* $Hash$*.* $\mathcal{B}$ *maintains a list* H_{list} *of tuples* (R, V, t, k, Q_{ca}) *which is initially empty, and a query counter* μ *which is initially set to* 0*.* $\mathcal{A}$ *provides a new pair* (R, V, t, Q_{ca}) *for hash query by first choosing a message* R *and then computing* $V = R + k * G$*, where* k *is a random number in* Z^*_n*. Upon a hash query* (R, V, t, Q_{ca}) *for which there exists a record* (R, V, t, Q_{ca}, h) *in* H_{list}*,* $\mathcal{B}$ *return* h *to* $\mathcal{A}$*; otherwise,* $\mathcal{B}$ *uniformly chooses a random number* $h \in Z^*_n$ *as the value of* $H(V, t, Q_{ca})$*, places* (R, V, t, Q_{ca}, h) *into* H_{list}*, and returns* h *to* $\mathcal{A}$*.*

Signature Queries:*Proceeding adaptively, the adversary* $\mathcal{B}$ *answers* $\mathcal{A}$*'s queries for signing oracle* $\mathcal{S}(d_{ca}, \cdot)$*. When* $\mathcal{A}$ *provides a query message* R*,* $\mathcal{B}$ *works as follows:*

(1) *Randomly choose two numbers* $u, v \in Z^*_q$*, and compute* $V = R + u * G + (-v * Q_{ca})$*;*
(2) *Set* $s = u$*,* $H(V, t, Q_{ca}) = v$*, and place* (R, V, t, Q_{ca}, v) *into* H_{list}*;*
(3) *Returns* (v, s) *as a signature on message* R *to* $\mathcal{A}$*.*

After $\mathcal{A}$ *makes* H*-hash query on* (R, V, t, Q_{ca}) *to get* $v = H(V, t, Q_{ca})$*, it can verify that* (V, s) *satisfies* $R + s * G = V + v * Q$*. Therefore,* (V, s) *is a valid signature on message* R *with respect to* j*'s public key* Q_{ca}*. Since* V *and* s *follows the uniform distribution, and* $Hash$ *behaves as a random oracle,* $\mathcal{A}$ *cannot distinguish between* $\mathcal{B}$*'s response and the real life.*

Output: *Eventually, suppose* $\mathcal{A}$ *returns a forgery* (R^*, V^*, t^*, s^*)*, where* (V^*, t^*, s^*) *is a valid forgery distinct from any previously given signature on message* R^* *with respect to the public key* Q_{ca}*.*

According to the above proof, $\mathcal{A}$ *can find a valid signature with non-negligible probability* $Adv^{cma}_{DS,\mathcal{A}}(\kappa)$*. Then, by using the forming lemma,* $\mathcal{A}$ *can outputs a new forgery* $(V^*, t^*, \hat{s}^*)$ *on the same message* R^* *and a different oracle* $\hat{Hash}(\cdot)$*, with non-negligible probability, such that* $Hash(V^*, t^*, Q^*_{ca}) \neq \hat{Hash}(V^*, t^*, Q^*_{ca})$ *and* $s^* \neq \hat{s}^*$*. From this, we get*

$$s^* = k + d_{ca} Hash(V^*, t^*, Q^*_{ca}) \bmod n$$
$$\hat{s}^* = k + d_{ca} \hat{Hash}(V^*, t^*, Q^*_{ca}) \bmod n$$

Thus, $\mathcal{B}$ *can solve out the private key*

$$d_{ca} = \frac{s^* - \hat{s}^*}{Hash(V^*, t^*, Q^*_{ca}) - \hat{Hash}(V^*, t^*, Q^*_{ca})} \bmod n \quad (2)$$

which is just $\mathcal{B}$*'s challenge* d_{ca}*. The choices of* i *in algorithm* $\mathcal{B}$ *imply that with probability at least* $\frac{1}{T_1(\kappa)}$ *he can 'hit' the correct value of* $\mathcal{A}$*. Thus,* $Adv^{ecdl}_{\mathcal{G},\mathcal{B}}(\kappa) \geq \frac{Adv^{cma}_{DS,\mathcal{A}}(k)}{T_1(\kappa)}$*, where* $Adv^{ecdl}_{\mathcal{G},\mathcal{B}}(\kappa) = Pr[pas \leftarrow Gen(1^\kappa); Q_{ca} \leftarrow \mathcal{G}; d_{ca} \leftarrow \mathcal{B}(pas, Q_{ca}) : d_{ca} * G = Q_{ca}]$*. Since the ECDLP is assumed to be hard in* $\mathcal{G}$*, then* $Adv^{ecdl}_{\mathcal{G},\mathcal{B}}(\kappa)$ *must be negligible. This contradicts the assumption that* $Adv^{cma}_{DS,\mathcal{A}}(\kappa)$ *is nonnegligible. Thus, we conclude that* $Adv^{cma}_{DS,\mathcal{A}}(\kappa)$ *is negligible for all adversaries* $\mathcal{A}$*.*

Theorem 2. *Assuming* (1) *ECDLP is hard for* $\mathcal{G}$ *and* (2) $Hash$ *is a random oracle, there is no adversary* A *that is successful against the proposed scheme SEIC.*

Theorem 3. *Assuming* (1) *ECDLP is hard for* $\mathcal{G}$ *and* (2) $Hash$ *is a random oracle, there is no adversary* A *that is successful against the proposed scheme SEIC.*

Proof 3. *Using Lemma 1, the proof can be obtained similar to the proof of Theorem 2 in [7]. So, we shall omit this proof.*